AMERICAN JOURNAL of PHYSICS

(Formerly THE AMERICAN PHYSICS TEACHER)

A Journal Devoted to the Instructional and Cultural Aspects of Physical Science

VOLUME 12, NUMBER 5 October, 1944

Atomic and Molecular Theory Since Bohr: Logical and Mathematical Survey

HENRY MARGENAU AND ARTHUR WIGHTMAN*

Yale University, New Haven, Connecticut

LOGICALLY, every physical theory presents two aspects, the inductive and the deductive. The first is one of particularized investigations leading to single discoveries until their pattern somehow emerges and impresses itself on the scientist as an embracive picture of reality. The other is an inspired guess at fundamental matters, a grand conjecture transcending the limits of experimentation, from which, by logical deductions, specific theorems flow. If these theorems are empirically verified, the initial conjecture becomes the basis of a valid theory. Philosophically, it is interesting to note that the empirical aspect of a science is never able to confer certainty upon scientific predictions, while the deductive never opens itself to complete proof. The predictive potency, the proper claim to fundamental understanding, are imparted to science by a remarkable interplay between inductive and deductive procedures.

The historical development of a science is rather neutral to this distinction; it moves into the foreground of attention sometimes the inductive, sometimes the deductive aspects of theory; more frequently it mixes them thoroughly. In the first paper of this series[1] the approach to recent atomic physics was chiefly historical. A fuller understanding will be gained if we supplement that treatment by a deductive, or axiomatic, presentation which better displays the coherence of knowledge recently gleaned. This mode of treatment has the further advantages of precision of thought and economy of statement, although it may be more abstract and therefore, perhaps, less satisfying to the experimental physicist than the inductive procedure.

In an appraisal of any theory, it is most important to note its limitations and its failures, for strangely they are the signposts pointing to the future. Only the axiomatic approach allows their exposition. It will be our endeavor in the following sections to review not only the successes which quantum theory has achieved in its many applications to atomic and molecular problems, but also its interesting shortcomings in the fields of relativity and radiation.

At the present stage of development, elegance of presentation and mathematical rigor seem to exclude each other. Dirac has achieved the former by taking his readers through a delightful mathematical fairyland; von Neumann has supplied much rigor at the expense of frightening many of his readers. In the present account we shall not be too preoccupied with mathematical questions, although some reference must be made to the exact treatment of von Neumann.

* At present in U. S. Navy.

[1] H. Margenau and A. Wightman, Am. J. Phys. 12, 119 (1944).

1. PHYSICAL STATES AND OBSERVABLES

In reference 1 there was some discussion of the need for a change in the classical concept of the state of a physical system, this need arising primarily from the introduction of probabilities which was enforced by the failure of older theories. Investigations mentioned in that article revealed that the probabilities themselves were not as important as the *probability amplitudes*. These are complex functions so defined that the squares of their absolute values are probabilities; for example, if $w_E(\mathbf{r})\mathrm{d}x\mathrm{d}y\mathrm{d}z$ is the probability that a particle with energy E is in the volume $\mathrm{d}x\mathrm{d}y\mathrm{d}z$ at the position $\mathbf{r}$, then $|\psi_E(\mathbf{r})|^2=w_E(\mathbf{r})$, where $\psi_E(\mathbf{r})$ is the probability amplitude (with respect to position). The laws which this amplitude has to obey were found by Schrödinger to be analogous to those that governed waves in certain classical problems. Like wave amplitudes, they are subject to the principle of superposition, for the addition of two probability amplitudes gives a new probability amplitude. This explains their name.

When such an amplitude is given, one can deduce the probabilities of a given set of results of *any* set of measurements on the physical system. It determines, therefore, all that one can possibly find out about the physical system. The probability of given results of measurement could be different for two systems only if their probability amplitudes were different. Thus the condition, or *state*, of a physical system seems to be characterized by its probability amplitude. This idea is contained in the first axiom of the theory:

Axiom 1. The states of a physical system may be represented by complex-valued functions $\psi(q)$ satisfying the condition

$$\int |\psi(q)|^2\mathrm{d}\tau \qquad (1)$$

exists and is finite.

As yet the arguments q of $\psi(q)$ and the volume element $\mathrm{d}\tau$ have not been specified. These will of course depend on the physical system under consideration.

In the important case of a particle one might expect the arguments of $\psi(q)$ to be the coordinates x, y, z of the particle. However, numerous observations show that elementary particles such as the electron, proton and neutron possess another degree of freedom, so that a fourth variable, somehow representing the *spin*, must be introduced. This variable is not continuous; the measurement of the spin of an electron is capable of only two independent results. For this reason the postulate has been introduced by Pauli that the spin variable s is to be regarded as capable of taking on two values only, $+1$ and -1. The integral (1), for the case of a single particle, must then be interpreted as a sum over the two values of s and an integral over the ordinary volume element; that is,

$$\sum_s \int |\psi(x, y, z, s)|^2\mathrm{d}x\mathrm{d}y\mathrm{d}z.$$

In the case of systems that are not particles, or systems containing more than one particle, q in Axiom 1 may stand for fewer or more than four variables. What they are is to be discovered in each case. Usually, they turn out to be the classical degrees of freedom.

If ψ satisfies Axiom 1, it will still do so after multiplication by a constant. This constant may be so chosen that $\int |\psi|^2\mathrm{d}\tau=1$. With this choice of constant, ψ is said to be *normalized*. Henceforth, all states will be assumed to be normalized.

It is well to note here that Axiom 1 can be replaced by a more demanding but at the same time more general requirement: *The states of a physical system shall be represented by points of a Hilbert space.*[2] This form of the axiom includes the matrix-mechanical and wave-mechanical definitions of state as special cases.

A *Hilbert space*, for which mathematicians use the symbol R_∞, is a set of abstract elements $\varphi, \psi, \cdots$ having the following properties.

I. *The elements form a "linear vector space."* If φ, ψ, χ are elements of R_∞ and a, b, c are complex constants, there are defined an addition and a scalar multiplication with the properties

$$\begin{gathered}\varphi+\psi=\psi+\varphi,\\ \varphi+(\psi+\chi)=(\varphi+\psi)+\chi,\\ a(\varphi+\psi)=a\varphi+a\psi,\\ (a+b)\varphi=a\varphi+b\varphi,\\ (ab)\varphi=a^*(b\varphi),\\ 0\cdot\varphi=0,\quad 1\cdot\varphi=\varphi.\end{gathered}$$

[2] See von Neumann, *Die mathematischen Grundlagen der Quantenmechanik* (Springer, 1932), pp. 18–101.

II. *The elements permit formation of a "Hermitean inner product."* For any pair of elements of R_∞, say φ and ψ, there is defined a certain complex number, denoted by (φ, ψ), called their *Hermitean inner product*, with the properties

$$(\varphi+\psi, \chi)=(\varphi, \chi)+(\psi, \chi),$$
$$(a\cdot\varphi, \psi)=a^*(\varphi, \psi),$$
$$(\varphi, \psi)=(\psi, \varphi)^*,$$
$$(\varphi, \varphi)\geq 0.$$

III. *The space has infinite dimensionality.* There exist arbitrarily many linearly independent elements of R_∞.

IV. R_∞ *is complete.* If the sequence $\{\varphi_i\}$ is such that

$$(\varphi_j-\varphi_k,\ \varphi_j-\varphi_k)$$

gets arbitrarily small when $j,\ k\geq N$, N being sufficiently large, then the sequence $\{\varphi_i\}$ possesses a *limit* φ in R_∞ such that

$$(\varphi_j-\varphi,\ \varphi_j-\varphi)$$

is arbitrarily small for a sufficiently large value of j.

V. *The space is separable.* There exists a sequence $\{\varphi_i\}$ in R_∞ such that the difference between any element ψ of R_∞ and some elements of $\{\varphi_i\}$ is arbitrarily small in the sense that $(\varphi_i-\psi,\ \varphi_i-\psi)$ vanishes.

From the preceding postulates one may deduce all the general theorems on Hilbert space. The relevance of these matters to our present purpose lies in the fact that *the states defined by $\psi(q)$ in connection with Axiom 1 are elements in Hilbert space.* They satisfy the requirements listed as properties of a linear vector space. They permit formation of a Hermitean inner product: If ψ and φ are states, and if their Hermitean product is defined as $\int\varphi^*(q)\psi(q)\mathrm{d}\tau\equiv(\varphi, \psi)$, all properties specified under Postulate II are reproduced. Also the last three postulates defining the properties of R_∞ may be shown to be satisfied by our probability amplitudes. From the technical point of view, therefore, the quantum-mechanical state has revealed itself as an inhabitant of R_∞, a new world recently explored by mathematicians.

A study of Hilbert space leads to a theorem—the Riesz-Fischer theorem—which may be stated, somewhat crudely, in this way: There exists an *isomorphism*, that is, a complete one-to-one correspondence almost amounting to identity, between all φ satisfying $\int|\varphi|^2\mathrm{d}\tau<\infty$ on the one hand, and the set of all sequences $\{x_1x_2\cdots x_n\cdots\}$ such that $\sum_i|x_i|^2<\infty$, on the other. The isomorphism is expressed by the correspondence

$$x_i\leftrightarrow\int\psi_i^*\varphi\mathrm{d}\tau\equiv(\psi_i, \varphi),$$

where $\psi_1, \psi_2\cdots$ is any complete set of orthonormal functions. In other words, for every state φ we can also define a *vector*, or column matrix, having components $x_i=\int\psi_i^*\varphi\mathrm{d}\tau$.

It is this isomorphism which was discovered by physicists when they developed two forms of quantum mechanics, that in terms of wave functions or probability amplitudes (Schrödinger) and that in terms of matrices (Heisenberg), which to their initial surprise yielded identical results.

In the following, we shall use predominantly the Schrödinger representation. General Hilbert space will occasionally be employed when this seems more convenient, chiefly in the present section. Familiarity with it will not be presupposed in the following sections.

Having characterized the states of a physical system we now proceed to establish a connection between the states of a physical system and the values of observable quantities measured on the system. First, however, we must make some mathematical definitions.

An *operator* is a correspondence between one point or element of Hilbert space and another, or, more simply, an operator converts one function into another. For example, $x\cdot$ and $\mathrm{d}/\mathrm{d}x$ are operators:

$$x\cdot f(x)=h(x),\quad \frac{\mathrm{d}}{\mathrm{d}x}f(x)=g(x).$$

If

$$\boldsymbol{R}f(q)=rf(q), \tag{2}$$

where r is a complex constant, then r is said to be an *eigenvalue* and f an *eigenfunction* of the operator $\boldsymbol{R}$. An operator $\boldsymbol{H}$ is called a *Hermitean operator* if, for any f and g in R_∞, $(\boldsymbol{H}f, g)=(f, \boldsymbol{H}g)$ or, in Schrödinger's representation,

$$\int(\boldsymbol{H}f)^*g\mathrm{d}\tau=\int f^*(\boldsymbol{H}g)\mathrm{d}\tau. \tag{3}$$

We are now ready to discuss Axiom 2.

Axiom 2. To each observable quantity there corresponds an Hermitean operator. The eigenvalues of the operator are possible values of the measured quantity; the eigenfunctions corresponding to these given eigenvalues represent states in which the physical system has exactly those eigenvalues.

Axiom 2 insures that only *real* observed values will be predicted by the theory, because the eigenvalues of a Hermitean operator are real: for if $\boldsymbol{H}f=\lambda f$, then $(\boldsymbol{H}f, f)=(\lambda f, f)=(f, \boldsymbol{H}f)=(f, \lambda f)=(\lambda f, f)^*$. Therefore $\lambda=\lambda^*$, which means that λ is real. A study of Hermitean operators reveals that usually not every real number is an eigenvalue, that is, the operator may have isolated real eigenvalues. All such are called the *point spectrum of the operator*. In case the operator has intervals of eigenvalues these are said to be its *continuous spectrum*.

The eigenfunctions belonging to Hermitean operators can all be made *orthogonal*. A set of functions $\varphi_1, \varphi_2, \varphi_3 \cdots$ is said to be orthogonal if, in the Schrödinger scheme,

$$\int \varphi_i^* \varphi_j d\tau = 0 \text{ when } i \neq j$$

or, if the φ_i are elements of Hilbert space, when

$$(\varphi_i, \varphi_j)=0, \quad i \neq j.$$

They are said to be *orthonormal* if, in addition, $(\varphi_i, \varphi_i)=1$. They are *complete* if it is possible to expand *any* φ, not a member of the set, in terms of the set, $\varphi=\sum_i b_i \varphi_i$ with b's that are complex constants. Eigenfunctions of Hermitean operators are in general orthonormal and complete.

The special choice of operators made in the application of the theory can only be justified empirically. In the Schrödinger theory one makes the correspondence (we write here $\hbar$ for $h/2\pi$, h being the Planck constant)

$$\left.\begin{array}{l} \text{momentum:} \\ \qquad p_i \rightarrow -i\hbar\partial/\partial q_i, \quad \text{or} \quad \mathbf{p} \rightarrow -i\hbar\nabla. \\ \text{energy:} \\ H(p, q) \rightarrow H(-i\hbar\nabla, q); \quad \text{also}^3 \quad E \rightarrow i\hbar\partial/\partial t. \\ \text{position coordinate:} \\ \qquad q_i \rightarrow q_i. \end{array}\right\} \quad (4)$$

[3] For an application of this operator E, see SEC. 2.

Then, for example, the eigenvalue equation for linear momentum (in one dimension) reads

$$-i\hbar \frac{\mathrm{d}}{\mathrm{d}q}\psi(q)=\lambda\psi(q).$$

Solving, we have $\psi(q)=c \exp(i\lambda q/\hbar)$. Here the condition that

$$\int_{-\infty}^{\infty} |\psi|^2 \mathrm{d}q < \infty$$

imposes no restrictions at all on the momentum, though it requires c to be an infinitesimal quantity. Hence any value of momentum is possible. This is not the case in the eigenvalue equation for the energy of a free particle which is confined to a length of x axis from 0 to L. Here the energy operator is

$$\boldsymbol{H}=\frac{p^2}{2m}=-\frac{\hbar^2}{2m}\nabla^2=-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2},$$

and the eigenvalue equation $\boldsymbol{H}\psi=E\psi$ has the solution

$$\psi=A \sin\sqrt{\left(\frac{2mE}{\hbar^2}\right)}x+B \cos\sqrt{\left(\frac{2mE}{\hbar^2}\right)}x.$$

But if $|\psi|^2$, and hence ψ, is to be zero at $x=0$ and at $x=L$, then B must be zero and $\sqrt{(2mE/\hbar^2)}L=n\pi$, where n is an integer. Hence $E=n^2\pi^2\hbar^2/2mL^2$; the energy is quantized. For particles bound by forces, $\boldsymbol{H}=-(\hbar^2/2m)\nabla^2+V(\mathbf{r})$, V being the potential energy, and the eigenvalue equation for the energy has a more complicated form. It is called the *Schrödinger equation* in all instances.

One further example which will be of interest later is the Schrödinger equation for the simple harmonic oscillator, for which $V=\frac{1}{2}m\omega^2x^2$. The solution of

$$\boldsymbol{H}\psi=\left(-\frac{\hbar^2}{2m}\frac{\mathrm{d}^2}{\mathrm{d}x^2}+\tfrac{1}{2}m\omega^2x^2\right)\psi=E\psi$$

leads to eigenfunctions known as *Hermite orthogonal functions*;[4] the energies are quantized and are given by

$$E=(n+\tfrac{1}{2})\hbar\omega, \quad (5)$$

with n a positive integer.

[4] See any book on quantum mechanics.

The choice of the operator $\mathbf{p}=-i\hbar\nabla$ can be justified in the following way. Suppose we wish to construct wave functions of the most general sort from plane waves $\exp i(\mathbf{k}\cdot\mathbf{r}-\omega t)$ for which the relations $\mathbf{p}=\hbar\mathbf{k}$, $E=\hbar\omega$ and an energy relation $E=E(p)$ hold. Let

$$\psi(\mathbf{r})=\int_{-\infty}^{\infty}\int\int\varphi(\mathbf{k})\exp i(\mathbf{k}\cdot\mathbf{r}-\omega t)\mathrm{d}k_1\mathrm{d}k_2\mathrm{d}k_3.$$

Then if we choose the nonrelativistic relation $E=p^2/2m$, we find that ψ is a general solution of

$$\left(-\frac{\hbar^2}{2m}\nabla^2-i\hbar\frac{\partial}{\partial t}\right)\psi=0,$$

an equation which will appear again in the discussion of the time dependence of states (SEC. 2). If one assumes the relativistic relation

$$E^2/c^2=m^2c^2+p^2,$$

then ψ is the general solution of

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}+\frac{m^2c^2}{\hbar^2}\right)\psi=0. \tag{6}$$

The essential fact is that both equations can be obtained by substituting

$$\mathbf{p}\rightarrow-i\hbar\nabla,$$
$$E\rightarrow i\hbar\partial/\partial t$$

in the energy relation $E=E(p)$.

Axiom 3. If a system is in a state φ, the probability of observing the eigenvalue r_i of the observable with operator $\boldsymbol{R}$ is $|(\varphi_i,\varphi)|^2$, where φ_i is the eigenfunction corresponding to the eigenvalue r_i.

The axiom implies that if φ be expanded as a series in the eigenfunctions of $\boldsymbol{R}$,

$$\varphi=\sum_j b_j\varphi_j,$$

then the probability that the eigenvalue r_i will be observed in a measurement of the observable R is given by the square of the absolute value of the coefficient b_i of φ_i. For

$$(\varphi_i,\varphi)=(\varphi_i,\sum_j b_j\varphi_j)=\sum_j b_j(\varphi_i,\varphi_j)=b_i,$$
$$|(\varphi_i,\varphi)|^2=|b_i|^2.$$

Further, $\sum_i|(\varphi_i,\varphi)|^2=(\varphi,\varphi)=1$, which is in accord with our probability interpretation. If $\varphi=\varphi_i$, then the measurement will certainly yield r_i.

The foregoing axiom has to be applied with care in two special cases. The first is that of degeneracy—the case in which two or more distinct eigenfunctions have the same eigenvalue. Then $|(\varphi_i,\varphi)|^2$ must be replaced by $\sum_\lambda|(\varphi_\lambda,\varphi)|^2$, in which the sum is taken over all states belonging to the given eigenvalue.

The second case requiring attention is that in which the eigenvalue, say E, belongs to a continuous spectrum. Here the usual method of the probability calculus must be applied. Since for a continuous spectrum the probability that E lies between E' and $E'+\Delta E'$ is infinitely greater than the probability that E has exactly the value $E'+\frac{1}{2}\Delta E'$, the probability per unit variation of eigenvalue must be introduced. This quantity can be secured by the construction of an eigendifferential

$$\psi(E',E'+\Delta E')=\int_{E'}^{E'+\Delta E'}\varphi_E\mathrm{d}E,$$

which may be normalized in the same way that the eigenfunctions of the point spectrum are normalized. Axiom 3 may now be used with $\psi(E',E'+\Delta E')$ in place of φ_i:

$$|(\psi(E',E'+\Delta E'),\varphi)|^2$$

is the probability that a measurement on a system in state φ shall yield a result E between E' and $E'+\Delta E'$.

It has here been tacitly assumed that the eigenfunctions of the operators of quantum mechanics form complete sets in the sense that no function can be orthogonal to all of them, and therefore that any function can be expanded in terms of them. This is one of the deepest mathematical questions of the operator theory and for its discussion we refer to other sources.[5] However, it is true that for any *state* function, that is, for any function such that $\int|\varphi|^2\mathrm{d}\tau$ converges,

$$\varphi=\sum_i b_i\varphi_i+\int b(E)\varphi_E\mathrm{d}E,$$

where φ_i are the eigenfunctions of the point spectrum of an Hermitean operator, and φ_E are the eigenfunctions of its continuous spectrum.

Another consequence of Axiom 3 is the "mean value theorem" which states: The expected mean of a large number of measurements on the observable corresponding to an operator $\boldsymbol{R}$, performed on a system in state φ, is given by $(\varphi,\boldsymbol{R}\varphi)$ or, in Schrödinger's representation, by $\int\varphi^*\boldsymbol{R}\varphi\mathrm{d}\tau$. This may be seen as follows (we omit for brevity all states of the continuous spectrum):

$$(\varphi,\boldsymbol{R}\varphi)=\sum_i(b_i\varphi_i,\boldsymbol{R}b_j\varphi_j),$$

where the φ_i are eigenstates of $\boldsymbol{R}$. The last expression is therefore

$$(\sum_i b_i\varphi_i,\sum_j b_jr_j\varphi_j)=\sum_i|b_i|^2r_i.$$

Now the r_i are eigenvalues, that is, possible results of the measurements, and $|b_i|^2$ represents the probability of their occurrence in conformity with our axiom. This proves the theorem. It is

[5] von Neumann, reference 2.

also possible to consider this theorem as the third axiom and to derive what we have called Axiom 3 from it.

2. MEANING OF MEASUREMENT; UNCERTAINTY; TIME DEPENDENCE OF STATES

It is clear from the foregoing discussion that the state of an atomic system, because of its postulated (and verified) relation to probabilities, has reference not to a single measurement but to an aggregate of measurements. When a state is completely known, only the probability distribution of the outcomes of all possible measurements is implied.

Conversely, a state cannot be determined by means of a single measurement. Failure to realize this obvious consequence of the probability situation has led to a great deal of loose discussion with regard to the philosophic foundations of quantum mechanics. If modern theory is correct, then the state of an atomic system bears the same relation (or lack of relation) to a single measurement as temperature does to the energy of a single molecule. A state cannot be determined by means of a single observation any more than the temperature of a gas can be specified when the kinetic energy of one of its molecules is known. This restriction in the meaning of a state represents a major departure from classical physics.

While there is no relation whatever between the results of two specific measurements performed on a system in a given state φ, the probability distributions of two sets of measurements of different observables are correlated. The most interesting and useful correlation of this type is the Heisenberg uncertainty principle; it has had a profound influence on the historical development of modern physics. It may be deduced at once from the axioms of SEC. 1.

One may show that if two Hermitean operators commute, then a system of orthonormal functions can be constructed which is a set of eigenfunctions for both. Thus if a system is in an eigenstate with respect to one observable, it will also be in an eigenstate with respect to the other: The corresponding two types of measurement are said to be compatible.

However, for noncommuting operators the situation is quite different. The noncommuting property reflects the physical situation in which the measurement of one physical quantity affects the possibility of measuring the other.

Consider two Hermitean operators $\boldsymbol{A}$ and $\boldsymbol{B}$ satisfying $\boldsymbol{AB}-\boldsymbol{BA}=c$, where c is a complex number. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are Hermitean, then it may be shown that c must be a pure imaginary number, which we shall put equal to $+ia$, where a is a real number. We choose as measure of the statistical spread of the measurements $\langle(A-\bar{A})^2\rangle_{Av}$, that is, the mean square deviation from the average of $\boldsymbol{A}$. Now

$$\begin{aligned}2\,\text{Imag}\,(\boldsymbol{A}\varphi,\boldsymbol{B}\varphi)&=i\{(\boldsymbol{A}\varphi,\boldsymbol{B}\varphi)-(\boldsymbol{B}\varphi,\boldsymbol{A}\varphi)\}\\&=-i\{(\boldsymbol{BA}\varphi,\varphi)-(\boldsymbol{AB}\varphi,\varphi)\}\\&=i(\{\boldsymbol{AB}-\boldsymbol{BA}\}\varphi,\varphi)=a(\varphi,\varphi)=a.\end{aligned}$$

Hence

$$\begin{aligned}1=\frac{2}{a}\,\text{Imag}\,(\boldsymbol{A}\varphi,\boldsymbol{B}\varphi)&\leqslant\frac{2}{|a|}|(\boldsymbol{A}\varphi,B\varphi)|\\&\leqslant\frac{2}{|a|}(\boldsymbol{A}\varphi,\boldsymbol{A}\varphi)(\boldsymbol{B}\varphi,\boldsymbol{B}\varphi),\end{aligned}$$

the last step by virtue of the Schwarz inequality.

Now $(\boldsymbol{A}-\bar{A})(\boldsymbol{B}-\bar{B})-(\boldsymbol{B}-\bar{B})(\boldsymbol{A}-\bar{A})=\boldsymbol{AB}-\boldsymbol{BA}=c$, since $\bar{A}=(\boldsymbol{A}\varphi,\varphi)$ and $\bar{B}=(\boldsymbol{B}\varphi,\varphi)$ are simply numbers. In other words, $\boldsymbol{A}-\bar{A}$ and $\boldsymbol{B}-\bar{B}$ satisfy the same commutation rule as the operators $\boldsymbol{A}$ and $\boldsymbol{B}$ themselves and may therefore be substituted into the last relation, which then reads

$$[\langle(\boldsymbol{A}-\bar{A})^2\rangle_{Av}]^{\frac{1}{2}}\cdot[\langle(\boldsymbol{B}-\bar{B})^2\rangle_{Av}]^{\frac{1}{2}}\geq\tfrac{1}{2}a. \qquad (1)$$

This is the uncertainty principle. It states that the spread of the measurements on A varies inversely as the spread of the measurements on B. If $\boldsymbol{A}$ and $\boldsymbol{B}$ are momentum and position operators relative to a particle, the constant a is equal to $\hbar$, as may be seen from the explicit form of these operators discussed in SEC. 1. The uncertainty relation therefore has its roots in the fact that the Planck constant has a finite value, that is, in the very phenomenon of quantization.

Thus far, no allowance has been made for any possible time variation in the state function φ; we have restricted ourselves to *stationary* states. What has been said about measurements referred to measurements made at the same instant, or at different instants provided these instants terminate equal periods after the preparation of the

state. In general, φ will change in time. When that is true, the constant E in the Schrödinger equation must be replaced by $i\hbar\partial/\partial t$, as was already noted in SEC. 1. The equation then reads

$$Hu = i\hbar\partial u/\partial t, \tag{2}$$

where we have written u in place of ψ to indicate the time dependence. This equation should be equivalent to

$$H\psi = E\psi \tag{3}$$

when H does not involve the time, that is, when the system is stationary. It is indeed; for Eq. (2) reduces to Eq. (3) when we put $u = \psi \exp(-iEt/\hbar)$. Since all that matters in the statistical interpretation is u^*u, and this is equal to $\psi^*\psi$, description in terms of u is equivalent to that in terms of ψ so long as states are stationary. But if the operator H changes its form in time, Eq. (2) regulates the system's behavior, and Eq. (3) is no longer useful.

An interesting application of the Schrödinger "time equation" (2) occurs in the study of the motion of a free particle. If it is assumed that at the present instant the probability of finding such a particle at x is given by

$$u^*u = ce^{-x^2/a^2}, \tag{4}$$

a Gauss distribution, then Eq. (2) will convert this probability into

$$u^*u = c\left[1+\left(\frac{\hbar}{ma^2}t\right)^2\right]^{-\frac{1}{2}} \exp\left\{\frac{-x^2}{a^2+\dfrac{\hbar^2}{m^2a^2}t^2}\right\} \tag{5}$$

at a later time t. The "wave packet," Eq. (4), has spread out into a *larger* packet (5). This phenomenon of diffusion is characteristic of all localized particles in quantum mechanics. If m is small, as for instance the mass of an electron, the rate of diffusion is very rapid; for ordinary masses it is imperceptibly slow.

Another application of Eq. (2) will be made in SEC. 5.

3. MANY PARTICLES; EXCLUSION PRINCIPLE

Thus far, states and observables of *single* systems have been under consideration. We shall now investigate how the states of systems containing more than one particle are to be described. Consider, first, two particles. It is clear that the variables to be used in constructing the state function of the pair are those characteristic of the first particle *and* those characteristic of the second.

In the simplest case we can say even more about the state function of the pair. Suppose that the particles are independent, or are isolated from each other, which is the physicist's way of saying that one is entirely uninfluenced by the other. Furthermore, for brevity we will ignore the spin. Assume that the probability of finding particle 1 at $x_1y_1z_1$ *in the absence of particle 2* is $w_1(x_1y_1z_1)$, and conversely that the probability of finding particle 2 at $x_2y_2z_2$ *in the absence of 1* is $w_2(x_2y_2z_2)$. Then it is a consequence of a theorem concerning the probability of the simultaneous occurrence of independent events that the probability of finding particle 1 at $x_1y_1z_1$ and at the same time particle 2 at $x_2y_2z_2$ is given by the product $w_1(x_1y_1z_1)\cdot w_2(x_2y_2z_2)$. If the reader now remembers the simple interpretation according to which $\varphi^2 = w$, it follows at once that the state function representing two *independent* particles is of the product form,

$$\varphi(x_1y_1z_1x_2y_2z_2) = \varphi_1(x_1y_1z_1)\cdot\varphi_2(x_2y_2z_2). \tag{1}$$

The same result may be derived mathematically from the axioms of SEC. 1. Let φ_1 be an eigenstate of the operator P_1 acting on the variables of the first system, p_1 being its eigenvalue. Then

$$P_1\varphi_1(x_1y_1z_1) = p_1\varphi_1(x_1y_1z_1).$$

Similarly, suppose that

$$P_2\varphi_2(x_2y_2z_2) = p_2\varphi_2(x_2y_2z_2).$$

If the individual systems are independent, the operator corresponding to the pair is $P = P_1 + P_2$; it is "separable" into summands each of which acts on only one set of coordinates. But the equation $P\varphi = p\varphi$ has the solution $\varphi = \varphi_1\cdot\varphi_2$ corresponding to the eigenvalue $p = p_1 + p_2$.

In general, when the state of one particle is influenced by the other, the coordinates $x_1\cdots z_2$ will appear in the function φ so thoroughly intermingled that their disentanglement into two factor functions is impossible. From the probability point of view the states of the individual particles are then intrinsically related. But even in that most general situation the function φ may

be expanded in terms of simple product functions; that is, it is possible to write

$$\varphi(x_1y_1z_1,\ x_2y_2z_2)=\sum_{ij} a_{ij}\varphi_i(x_1y_1z_1)\cdot\varphi_j(x_2y_2z_2) \quad (2)$$

with constant values of the a_{ij}. Only if all a_{ij} can be written in the form $\alpha_i\beta_j$ will the right-hand member of Eq. (2) represent a simple product, that is, describe independent systems.

An extension of this scheme to more than two particles at once suggests itself. If the state function of two independent particles is a product of two factors, the state function describing n independent particles is a product of n factor functions; thus,

$$\varphi(x_1\cdots z_n)=\varphi_1(x_1y_1z_1)\cdot\varphi_2(x_2y_2z_2)\cdots\varphi_n(x_ny_nz_n). \quad (3)$$

Here again, interaction between the particles manifests itself in a greater complexity of φ, its symptom being failure of the state function to be factorable in the manner of Eq. (3).

The results thus far considered are profoundly modified by the exclusion principle, understanding of which is best prepared by a discussion of the symmetry properties of a function of several variables.

Mathematical functions are classified in accordance with their properties relative to mathematical operations. For instance, to determine whether a function is even or odd, the operation "change of sign of the argument" is performed: if the function itself retains its value it is said to be *even;* if it changes its sign it is *odd*. In a similar way, through definite operations, arise the common properties of continuity, differentiability, integrability, orthogonality with respect to other functions, convergence, and so on. All the properties just mentioned refer to operations that can be carried out on functions having but a single variable and hence are most widely known. In dealing with functions of many variables, such as the state functions of many-particle systems, certain *symmetry* operations become of great importance; they have enjoyed but little interest in orthodox function theory though more in the theory of groups. One of these symmetry operations is called an "exchange operation," or a "particle exchange." It is simply an interchange of the set of coordinates of one particle with that of another. Thus, the exchange operation relative to particles 1 and 2 would transform the function displayed in Eq. (3) into

$$\varphi_1(x_2y_2z_2)\cdot\varphi_2(x_1y_1z_1)\cdots\varphi_n(x_ny_nz_n).$$

Since n particles can form $\frac{1}{2}n(n-1)$ pairs, this is also the number of exchange operations performable on a state function describing n particles.

Henceforth we shall simplify matters by writing a single letter for the entire set of coordinates of one particle: We abbreviate $(x_iy_iz_i)$ to (i). Equation (3) then reads $\varphi=\varphi_1(1)\cdot\varphi_2(2)\cdots\varphi_n(n)$. We shall also write P_{ij} for the operation transposing set (i) and set (j); that is, $P_{12}\varphi_1(1)\cdot\varphi_2(2)=\varphi_1(2)\varphi_2(1)$. Furthermore, let us return to the case of two particles.

The effect of an exchange operation upon a function of two sets of variables (1, 2) may obviously be stated in one of three ways:

(a) the function $P_{12}(1, 2)$ is some altogether different function;
(b) $P_{12}\varphi(1, 2)=\varphi(1, 2)$;
(c) $P_{12}\varphi(1, 2)=-\varphi(1, 2)$.

The usual effect is described by (a); but (b) and (c) represent interesting and most important exceptions to this indiscriminate behavior. A function satisfying proposition (b) is said to be "symmetrical with respect to particle exchange," or simply *symmetrical;* (c) defines an *antisymmetrical* function.

Now the Pauli exclusion principle says that *state functions representing several similar particles must be antisymmetrical.* This is the accurate statement of this principle, which was discovered by Heisenberg in 1926 after Pauli, in 1925, had formulated it in a language referring to the older quantum theory. Let us see how detrimental it is to our former conclusions, particularly those concerning Eqs. (1) and (2).

We saw that *independent* particles have state functions of the product form: $\varphi(1, 2)=\varphi_1(1)\cdot\varphi_2(2)$. If φ_1 and φ_2 are *different* functions, say u and v, that is, if the individual particles are in different states, then $P_{12}u(1)v(2)=u(2)v(1)\neq\pm u(1)v(2)$. The function is neither symmetric nor antisymmetric and does not satisfy the exclusion principle. How, then, can it be made antisymmetric? Mathematicians know that there

is only one way of "antisymmetrizing" a function like $u(1)v(2)$; this is to subtract from the function itself its exchanged form. Thus the antisymmetrical function corresponding to $u(1)v(2)$ is

$$\varphi_A = u(1)v(2) - P_{12}u(1)v(2) = u(1)v(2) - u(2)v(1). \quad (4)$$

This conclusion is inescapable in spite of the serious implication that the result has no longer the form of a single product! It is the difference of two products and therefore of the form of Eq. (2), which we have recognized as the manifestation of *interdependence* of the particles. The exclusion principle, by merely stipulating antisymmetry, automatically introduces correlations between the states of the two particles. Although the correlations are of nondynamical origin, arising as they do from a formal principle of symmetry, they have the same physical effects as if they were due to forces.

Next, assume the two states to be equal, $u=v$; then $\varphi_A=0$. From the meaning of $|\varphi_A|^2$ as a probability we may then conclude that the chance of finding either particle is zero for every point of space; in other words, such states do not exist. This circumstance is often expressed by saying: Two similar particles may not be in the same state.

If the number of particles exceeds two, the construction of an antisymmetric function from a product of type (3) proceeds as follows. By performing the $\frac{1}{2}n(n-1)$ exchange operations singly and in every combination upon the function (3), all possible distributions of arguments may be achieved among the n individual-particle functions φ_i. Each of these distributions is called a *permutation*. Some permutations are produced by an *even* number of exchanges, others by an *odd* number. The former are called *even* permutations, the latter *odd* ones. The reader can easily verify the following rule for antisymmetrizing function (3), or any other function. First, form all possible permutations. Then affix a + sign to the even ones, a − sign to the odd ones, and add them all together. The result, which contains $n!$ terms, is the antisymmetrical combination required by the exclusion principle. This process is unique.

Again, the resulting function, which is the only one acceptable in view of the exclusion principle, does not permit the interpretation of independence among the particles, for it is not of the simple product type. The states are correlated in a peculiar way. If *any two* individual particle functions in Eq. (3) are the same, the whole antisymmetric combination vanishes. This leads again to the conclusion that no two similar particles may be in the same state.

There is an interesting and far-reaching parallelism between the general principle of relativity and the exclusion principle. The former creates physically perceptible forces out of the metric of space; by endowing its equations with the formal property of invariance it is able to account for the phenomenon of gravitation, no reference being made to the ordinary concept of force. The exclusion principle imposes another formal property, antisymmetry, upon the state functions of quantum physics and thereby yields correlations that are tantamount to forces. The physicist in fact calls them *exchange forces* without any apparent embarrassment.

It is to be remembered that the exclusion principle places no requirement upon the state functions describing particles of different species. Thus in the theory of the deuteron, which consists of a proton and a neutron, it finds no application at all. The normal state of that system is in fact symmetric. Similarly, the state function of the helium nucleus, which contains two protons and two neutrons, need not be antisymmetric with respect to a proton-neutron exchange, but only to proton-proton and neutron-neutron exchange. It also follows that the state functions must be *symmetric* with respect to an interchange of *pairs* of elementary particles. But further pursuit of these considerations, which would lead to the interesting results of quantum statistics, is beyond the scope of this report.

It may easily be shown that a state function satisfying the exclusion principle leads to the same observable properties for one particle as for any other. The mean value of the observable P_1 (having reference to particle 1) is $\int \varphi^* P_1 \varphi d\tau_1 d\tau_2 \cdots d\tau_n$. If in this integral the variables of particles 1 and 2 are interchanged, its value is of course unaltered; but it may then also be written $\int (-\varphi^*) P_2 (-\varphi) d\tau_1 \cdots d\tau_n$, which on cancellation of the two minus signs is the mean

value of the observable P_2. Now it is a physical fact that elementary particles of the same species cannot be empirically distinguished. The exclusion principle is clearly in harmony with this conclusion. Whether it implies Leibnitz' principle of the identity of indiscernibles, as is sometimes thought, is a matter into which we shall not enter here.[6]

To show more clearly the effect of the correlations introduced into the motion of dynamically independent particles we present a simple example. A single particle which moves along the x direction with momentum $k\hbar$ is represented by a state function $u(1)=e^{ik_1x_1}$. If there is another similar particle, also constrained to move along x but with momentum $k_2\hbar$, its state function is $e^{ik_2x_2}$. Hence

$$\varphi_A=e^{i(k_1x_1+k_2x_2)}-e^{i(k_2x_1+k_1x_2)},$$

so that

$$|\varphi_A|^2=2[1-\cos(k_1-k_2)(x_1-x_2)]. \qquad (5)$$

Whatever the values of k_1 and k_2, this quantity is zero provided $x_1=x_2$: The two particles cannot be at the same place. But $|\varphi_A|^2$ is also zero when $k_1=k_2$, whatever the positions; hence the particles cannot have the same momentum. Equation (5) is interesting in one further respect, for it shows that the tendency of the particles to avoid each other is greater the more nearly their momentums are equal. If we denote by x' the value of x_1-x_2 for which the probability $|\varphi_A|^2$ is $\frac{1}{2}$, we find

$$x'=\text{const.}/(k_1-k_2).$$

Thus the range of spatial exclusion is greater the smaller the difference in momentum. Since the particles were considered free (compare their state functions!), the repulsive effect here encountered is not of dynamic origin.

The exclusion principle makes a composite system *more* than the sum of its parts. In this respect it is probably unique. In a rudimentary way, the principle may contain the first elements necessary for an understanding of biological organization. It is because of its potential fruitfulness in larger domains that we have here accorded it a somewhat elementary and lengthy exposition.

[6] For the manifold philosophic implications of the exclusion principle see H. Margenau, J. Phil. Sci., in press.

4. RELATIVITY AND QUANTUM MECHANICS

The restricted theory of relativity requires that the equations of physics retain their form in all inertial systems; they must be invariant when subjected to a Lorentz transformation. The preceding results do not possess this property.

Doubts have been voiced as to the possibility of achieving such invariance; for there appears to be a fundamental conflict in the methods of quantum mechanics on the one hand and relativity theory on the other. The former involves the uncertainty principle, which prohibits the velocity of a particle from being known when its position is exactly given. The Lorentz transformations, however, represent relations between exact coordinates and operate at the same time with exact velocities of reference systems. While at first sight this circumstance seems to involve a contradiction, the difficulty resolves itself when we note that the coordinates which occur as arguments in the state function of quantum mechanics need not be *observable* and sharp. The Schrödinger equation itself contains the operators $\mathfrak{p}$ and $\boldsymbol{x}$ and does not, of course, contradict the uncertainty principle. There is consequently no formal reason why quantum mechanics should not be amenable to relativistic treatment.

The range of problems to which relativistic quantum theory can be applied is limited by the difficulties inherent in classical relativity itself. As is well known, there is no way of dealing exactly with the two-body problem on the basis of relativistic dynamics. We should expect, therefore, that the quantum-mechanical formalism when developed relativistically will fail before the many-body problem. Similarly, the difficulties which attach to the treatment of a radiating charge in ordinary electrodynamics will be duplicated when that problem is treated from the point of view of relativistic quantum theory. In addition to these troubles there are others that could not have been predicted in view of the normal delinquencies of relativity, troubles that arise out of the fusion of the two theories. Chief among these is the emergence of states of negative kinetic energy to which attention will be called later.

The first attempts to join relativity and quantum mechanics were made by Klein and by

Gordon,[7] whose endeavor was to modify the Schrödinger equation by the addition of terms of first and higher powers of $1/c^2$ in order to achieve invariance. While this process cannot be unique, they succeeded in obtaining an equation that appeared plausible and attractive indeed. Putting

$$x_1=x, \quad x_2=y, \quad x_3=z, \quad x_4=ict,$$

$$\phi_1=\frac{e}{c}A_x, \quad \phi_2=\frac{e}{c}A_y, \quad \phi_3=\frac{e}{c}A_z, \quad \phi_4=\frac{e}{c}i\varphi,$$

where **A** is the vector potential and φ the scalar potential, they arrived at

$$(\sum_1^4 P_k^2+m^2c^2)u=0 \tag{1}$$

with

$$P_k=-i\hbar\frac{\partial}{\partial x_k}+\phi_k.$$

That this equation has proper relativistic form may roughly be seen from the symmetry with which space and time coordinates enter. It also reduces to the correct description for $c\to\infty$, as will now be shown.

If we put $\mathbf{A}=0$, write $e\varphi=V$, and $u=\psi\exp(i\epsilon t/\hbar)$, Eq. (1) becomes

$$\left(-\hbar^2\nabla^2-\frac{\epsilon^2}{c^2}+\frac{2\epsilon V}{c^2}+\frac{V^2}{c^2}+m^2c^2\right)\psi=0.$$

Now the relativistic energy ϵ is equal to mc^2+E, where E represents the ordinary energy appearing in the former treatment. On substituting this value of ϵ and retaining only terms free from $1/c^2$, we find that the last equation immediately takes the form

$$(-\hbar^2\nabla^2-2mE+2mV)\psi=0,$$

which is Schrödinger's equation.

For free particles ($\mathbf{A}$, $\varphi=0$), the Klein-Gordon equation becomes

$$\left[\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\left(\frac{mc}{\hbar}\right)^2\right]u=0. \tag{2}$$

Except for the last term, this is the wave equation representing the behavior of photons. It is often regarded as more general than the wave equation —as describing the action of *all* particles, material as well as photons. The latter are then to be considered as particles having a rest mass $m=0$.

This point of view has led to the "meson theories" of nuclear forces which are of considerable interest at present. Briefly, the idea is this. If m is put equal to zero in Eq. (2), the latter represents electromagnetic waves in vacuum. The generalization necessary in order to treat the case in which *sources of potential*—namely, electrical charges—are present is to replace the zero in the right-hand member by $-4\pi\rho$. The equation then reads

$$\left(\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)u=-4\pi\rho.$$

The solution of this equation is known from electrodynamics; it is

$$u=\int\rho'd\tau'/|\mathbf{r}-\mathbf{r}'|,$$

ρ' being the retarded charge density at the point $\mathbf{r}'$.

What would be the meaning of a similar generalization of Eq. (2) with m not made equal to zero? The suggestion is strong that the equation

$$\left[\nabla^2-\frac{1}{c^2}\frac{\partial^2}{\partial t^2}-\left(\frac{mc}{\hbar}\right)^2\right]u=-4\pi\rho \tag{3}$$

gives the potential u due to particles of mass m, in the presence of a *source of particles* of density ρ. Now there are reasons why protons and neutrons may perhaps be regarded as sources of mesons, the latter having a mass approximately 150 times the electronic mass, and it is therefore tempting to suppose that the solution u of Eq. (3) represents the potential of one source in the meson field due to another. But the solution of Eq. (3) may be shown to be

$$u=\int\frac{\rho'd\tau'}{|\mathbf{r}-\mathbf{r}'|}\left[\exp\left(-\frac{mc}{\hbar}|\mathbf{r}-\mathbf{r}'|\right)\right]$$

or, if the source has no spatial distribution, that

[7] O. Klein, Zeits. f. Physik **37**, 895 (1926); W. Gordon, Zeits. f. Physik **40**, 117 (1926).

is, is a point source,

$$u=\frac{\text{const.}}{r}\left[\exp\left(-\frac{mc}{\hbar}r\right)\right]. \tag{4}$$

The interesting result, noticed first by Yukawa,[8] is the fact that if m is set equal to the experimental meson mass (which is not known very accurately, however), then the potential u given by Eq. (4) has the range of action previously postulated for the potential between nuclear particles, $\hbar/mc\approx 2.5\times 10^{-13}$ cm. The meson theory asserts that this is not a coincidence.

When the Klein-Gordon equation is applied to problems in which $\mathbf{A}$ and φ are not zero, it leads, unfortunately, to erroneous results. It does not, for example, give a correct description of the fine structure of the hydrogen energy levels. Thus, though full of useful suggestions, it had to be abandoned in favor of a more thoroughgoing revision of the whole problem than was provided by the method of intuitive supplementation proposed by Klein and Gordon. The successful attack was made by Dirac.[9]

In discussing Dirac's treatment we shall first confine our attention to the case of a free particle. The method is essentially identical with the one that led to success in the nonrelativistic case: In writing the Hamiltonian energy function, certain variables appearing in it will be replaced by operators; but the correspondence between variables and operators is different in some details than it was in the Schrödinger theory.

The relativistic energy H of a free particle can be expressed in two ways:

$$H^2=c^2p^2+m^2c^4 \tag{5}$$

and

$$H=\mathbf{v}\cdot\mathbf{p}+\sqrt{(1-\beta^2)}mc^2, \tag{6}$$

where $\mathbf{p}$ and $\mathbf{v}$ are the momentum and velocity of the particle, and $\beta=v/c$. If we start with Eq. (5) and make the usual replacement for $\mathbf{p}$, we obtain the free-particle form of the Klein-Gordon equation, which we know to be unsatisfactory. Hence we are led to use Eq. (6).

There is another reason for preferring the *linear* expression for H. If we wish to retain the old operator association between H and $i\hbar\partial/\partial t$ which was so successful in describing the variation of nonrelativistic states in time, Eq. (5) yields a time equation of the *second* order in t, whereas Eq. (6) gives a first-order equation. But a second-order equation can be used to predict a state $u(t)$ only if $u(t_0)$ *and* $\partial u/\partial t|_{t_0}$ are given at some time t_0. In other words, the *state* of a system is not sufficient information for predicting a future state; the *time derivative* of the state is required also. This would either spoil the meaning of the concept *state*, or else require modification of the very foundations of quantum mechanics.

If Eq. (6) is to be transcribed into operator form, we have to dispose of the troublesome variable $\mathbf{v}$ which did not appear in the classical Hamiltonian. In classical theory, $\mathbf{v}$ and $\mathbf{p}$ are in constant ratio, so that $\mathbf{v}$ may be expressed in a trivial manner in terms of $\mathbf{p}$. In relativity, however, this is no longer true, and $\mathbf{v}$ is a dynamical variable in its own right. We must be prepared, therefore, to reserve for it a special vector operator which we call $c\boldsymbol{\alpha}$. It is written in this way to make $\boldsymbol{\alpha}$ dimensionless. Thus have been introduced *three* single operators—the space components of $\boldsymbol{\alpha}$—which will be denoted separately either by α_x, α_y, α_z, or by α_1, α_2, α_3, whichever is more convenient in the context. Another operator will be assigned to the quantity $\sqrt{(1-\beta^2)}$, for there is no certainty that this square root can be uniquely expressed in terms of the α's. Let us write α_4 for it. In a moment it will be seen that there are just enough relations available to make possible the definition of four new operators. It would simplify matters if these operators did not act on space coordinates and therefore commuted also with the components of $\mathbf{p}=-i\hbar\nabla$; hence we shall try this assumption. It will be seen to be successful.

On using these substitutions in Eq. (6) and writing $H\psi=\epsilon\psi$, we find that there results Dirac's equation for the free electron,

$$(-i\hbar c\boldsymbol{\alpha}\cdot\nabla+\alpha_4 mc^2)\psi=\epsilon\psi. \tag{7}$$

The function ψ must now depend on more variables than x, y, z alone, for it has to be susceptible of modification by the α-operators which do not act on space coordinates. The mystery of the missing coordinate will soon resolve itself.

[8] H. Yukawa, Proc. Phys. Math. Soc. Japan **17**, 48 (1935).

[9] P. A. M. Dirac, *Quantum mechanics* (Oxford Press, ed. 2, 1935).

The nature of the α's is thus far undetermined. Dirac was able to specify them by the ingenious postulate that the operator $\boldsymbol{H}$, when applied twice in succession, shall have the form (5). This requires, as can be seen by expanding $\boldsymbol{H}^2$, the following commutation law between the α-operators:

$$\alpha_i\alpha_j+\alpha_j\alpha_i=2\delta_{ij}, \quad i, j=1, 2, 3, 4. \tag{8}$$

What, then, is the precise form of the α's? Commutation laws of type (8) do not define uniquely any specific mathematical construct; they can be satisfied by differential operators—symmetry operators, matrices, and so on. The simplest procedure is to represent these entities as *matrices*. It is found possible to invent square four-rowed matrices (matrices of lower order will not do!) that satisfy (8); the reader will verify that the following is such a set:

$$\alpha_1=\begin{pmatrix}0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\end{pmatrix}, \quad \alpha_2=\begin{pmatrix}0&0&0&-i\\0&0&i&0\\0&-i&0&0\\i&0&0&0\end{pmatrix},$$

$$\alpha_3=\begin{pmatrix}0&0&1&0\\0&0&0&-1\\1&0&0&0\\0&-1&0&0\end{pmatrix}, \quad \alpha_4=\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&-1&0\\0&0&0&-1\end{pmatrix}.$$

The nature of ψ depends on the choice of α_i; if the latter are matrices, ψ must be a column vector, that is, a matrix having but one column:

$$\psi=\begin{pmatrix}\psi_1\\\psi_2\\\psi_3\\\psi_4\end{pmatrix}.$$

Each component is, of course, a function of x, y, z. The subscript on ψ may be regarded as a fourth variable capable of taking on only the discrete values 1, 2, 3, 4. In the following, however, we shall adhere to the matrix representation and attach indices to ψ when referring to its components.

Normalization of ψ now requires

$$\psi^\dagger\psi\equiv\sum_{i=1}^{4}\psi_i^*\psi_i=1,$$

and the mean-value relation for an operator $\boldsymbol{P}$ reads,

$$\bar{P}=\int\psi^\dagger\boldsymbol{P}\psi\,d\tau\equiv\int\sum_i\psi_i^*(\boldsymbol{P}\psi)_i\,d\tau. \tag{9}$$

To obtain the eigenvalues of Dirac's equation is a simple matter. If the electron is supposed to move along the x axis, the ψ_i will not depend on y and z. Equation (7) then takes the expanded form

$$(\epsilon-\alpha_4mc^2)\psi+i\hbar c\alpha_x\frac{d\psi}{dx}=0.$$

This set of four equations has solutions of the form $A\exp(ipx/\hbar)$, where A is another column vector with components A_1 to A_4. On inserting this the equation reduces to

$$(\epsilon-mc^2\alpha_4-pc\alpha_x)A=0, \tag{10}$$

which again represents a set of four equations. Instead of solving them explicitly, we "iterate" (apply twice in succession) the operator to the left of A. If then we use the commutation laws for the α's, the result is

$$(-\epsilon^2+m^2c^4+c^2p^2)A=0.$$

This is still a set of four equations, but the operator acting on A has lost its teeth: It contains no matrices and therefore merely multiplies all components of A. If at least one of these components is to be different from zero, $-\epsilon^2+m^2c^4+c^2p^2=0$, which requires

$$\epsilon=\pm\sqrt{(m^2c^4+c^2p^2)}. \tag{11}$$

This is, in fact, the ordinary result for the relativistic energy of a free particle, as may be seen at once from Eq. (5). Also, if we replace p by its classical equivalent $mv(1-\beta^2)^{-\frac{1}{2}}$, Eq. (11) reads $\epsilon=\pm mc^2(1-\beta^2)^{-\frac{1}{2}}$, which will be recognized as a familiar result.

In classical relativistic dynamics, the negative sign in Eq. (11) is ignored on the grounds that free particles with negative energy are never found in nature. This is proper, for there exists in that theory no mechanism which will change a particle with positive energy into one with negative energy; the division between possible and impossible states is a perfectly stable one.

This position cannot be maintained in quantum mechanics for two stringent reasons. In the first place, quantum jumps from positive to negative

energies *do* occur as further calculations show: The division becomes illusory. Second, if all states corresponding to negative energies are ignored, the mathematical scheme loses its significance, the set of functions generated by Dirac's equation is no longer complete and all expansions become invalid. Hence the negative energies must be retained. Their troublesome meaning will be discussed later.

Other interesting facts may be deduced from Eq. (7). For example, the solution $\psi = A \exp(ipx/\hbar)$ satisfies the equation

$$-i\hbar \frac{\mathrm{d}}{\mathrm{d}x}\psi = p\psi$$

and represents therefore an eigenstate not only of E but also of the linear momentum operator belonging to the eigenvalue p. Let us compute

$$\langle v_x \rangle_{Av} = c \int \psi^\dagger \alpha_x \psi \mathrm{d}\tau \Big/ \int \psi^\dagger \psi \mathrm{d}\tau = cA^\dagger \alpha_x A / A^\dagger A.$$

This may be evaluated by using Eq. (10), its associate, and the commutation laws for the α's, yielding

$$\langle v_x \rangle_{Av} = c^2 p/\epsilon, \tag{12}$$

a result that is reasonable enough. If, on the other hand, we compute $\langle v_x^2 \rangle_{Av}$, the result is

$$\langle v_x^2 \rangle_{Av} = c^2 \int \psi^\dagger \alpha_x^2 \psi \mathrm{d}\tau \Big/ \int \psi^\dagger \psi \mathrm{d}\tau = c^2 \tag{13}$$

because α_x^2 is 1 in view of Eq. (8). How can these two results, Eqs. (12) and (13), be reconciled? According to the first result, the electron progresses along the x axis with its classical speed c^2p/ϵ which, in view of the two unnumbered relations following Eq. (11), is nothing other than v_x itself. According to the second result, $\langle v_x^2 \rangle_{Av} > \langle v_x^2 \rangle_{Av}$. This can only mean that, besides moving along x, the electron moves rapidly in other directions also. Since, as may easily be seen, $\bar{v}_y = \bar{v}_z = 0$, this concurrent motion must be periodic about the x axis. It is to be interpreted as a "trembling" motion which accompanies the straight-line propagation. Further investigation[10] shows its amplitude to be very small indeed, of the order of magnitude $\hbar/mc = 3.8 \times 10^{-11}$ cm. This, together with other indications to be discussed in the next section, shows that the predictions of present relativistic quantum mechanics are not meaningful when they refer to phenomena taking place in very small regions of space.

Another peculiarity of the relativistic free-particle problem was exhibited by Klein and is often referred to as the "Klein paradox." It results when Eq. (7) is solved for two regions, one in which the potential energy V is zero and one in which it is finite and constant. On joining the two solutions and then computing the reflection coefficient due to the barrier V, one finds

$$R = \frac{4V^2m^2c^4}{[(P+P')^2 - V^2][(P+P'^*)^2 - V^2]},$$

where

$$P^2 = \epsilon^2 - m^2c^4, \quad P'^2 = (\epsilon - V)^2 - m^2c^4.$$

If P' is real, R is positive and, of course, smaller than 1; if P' is imaginary, simple calculation shows that $R = 1$. Hence, so long as P' is imaginary, all electrons are reflected by the barrier, otherwise they have a chance of going through. But the condition that P' be imaginary is $mc^2 > |\epsilon - V|$, and this means $\epsilon + mc^2 > V > \epsilon - mc^2$. In terms of the nonrelativistic energy $E = \epsilon - mc^2$, the condition reads $E + 2mc^2 > V > E$. Outside of this range for V, the barrier transmits. The *lower* limit of reflection is in accord with nonrelativistic theory, which also predicts certain reflection when $V > E$. The puzzling thing is the *upper* limit—the conclusion that *transmission* sets in again when $V > E + 2mc^2$. It may be shown that even for an infinitely high barrier the transmission coefficient is finite and takes on the value $2P/(\epsilon + P)$. Passage across the barrier is equivalent to conversion of the electron from a state of positive to one of negative energy, the energy in the barrier region being $\epsilon - V$ relative to the top of the barrier. If $V > \epsilon + mc^2$, this quantity is smaller than $-mc^2$. We see, then, that a simple barrier, if sufficiently high, can induce the troublesome transitions.

When first announced, the Klein paradox created doubt as to the validity of Dirac's theory. A little later, however, it developed that the steepness of the barrier has much to do with the

[10] E. Schrödinger, Sitz. Preuss. Akad. Wiss. **3**, 63 (1931).

anomaly. In fact, if the barrier does not rise as much as mc^2 energy units within a distance $\hbar/mc$, the paradox disappears. It probably is another instance of a false prediction concerning phenomena in very small domains of space.

The chief success of Dirac's theory results from its application to the hydrogen problem. To discuss this problem we must first generalize Eq. (7) so as to include forces that are derivable from a vector potential $\mathbf{A}$ and a scalar potential φ. In the presence of such force fields, Eq. (5) is known to take the form

$$(H+e\varphi)^2=(c\mathbf{p}+e\mathbf{A})^2+m^2c^4, \qquad (5')$$

while Eq. (6) can no longer be written in simple form. In spite of this, Eq. (5′) suggests that the desired generalization may be effected by replacing in Eq. (7) the operator $\mathbf{p}=-i\hbar\nabla$ by $-i\hbar\nabla+(e/c)\mathbf{A}$, and ϵ (or H) by $\epsilon+e\varphi$. This yields

$$H\psi=\left\{c\boldsymbol{\alpha}\cdot\left(-i\hbar\nabla+\frac{e}{c}\mathbf{A}\right)+\alpha_4mc^2-e\varphi\right\}\psi=\epsilon\psi \qquad (14)$$

as the general form of Dirac's equation.

In applying Eq. (14) to the hydrogen problem, $\mathbf{A}$ is put equal to zero, the α's are inserted in their matrix form, φ is made e/r and the four resulting first-order equations are solved. The resulting eigenvalue is

$$\epsilon=mc^2\left\{1+\frac{f^2}{([(j+\tfrac{1}{2})^2-f^2]^{\frac{1}{2}}+n_r)^2}\right\}^{-\frac{1}{2}}.$$

This is the famous fine-structure formula (see reference 1) which was first derived by Sommerfeld on the basis of Bohr's theory; it is known to be well substantiated by experiment.[11] Here j and n_r are quantum numbers, and $f=e^2/\hbar c=1/137$.

Although the ability of Dirac's theory to produce this result elegantly and without special assumptions as to spin properties of the electron is most noteworthy and indeed surprising, it represents but half of its triumph; for the theory also explains the *spin*. The easiest way to see this is the following.

[11] As to the intensities of the fine-structure components, the original Sommerfeld and the Dirac theories disagree. Experiment decides in favor of the latter.

In the nonrelativistic treatment of the hydrogen atom, the operator $\mathbf{L}=\mathbf{r}\times\mathbf{p}=-i\hbar\mathbf{r}\times\nabla$ commutes with the Hamiltonian; the orbital angular momentum of the electron is therefore a constant of the motion. If the reader will take the trouble to compute the commutation properties of the same operator with Dirac's H, that is, with

$$-ic\hbar\boldsymbol{\alpha}\cdot\nabla+\alpha_4mc^2-e\varphi(r),$$

he will find this to be no longer true. In fact, it is easily seen on the basis of Eq. (8) that

$$HL_z-L_zH=\hbar^2c\left(\alpha_x\frac{\partial}{\partial y}-\alpha_y\frac{\partial}{\partial x}\right).$$

Hence the electron's orbital angular momentum is *not* a constant of the motion. This can only mean that the orbital angular momentum is not the only component of angular momentum possessed by the electron, for if the total angular momentum were not constant, the electron would be subject to an external torque, which is contrary to the facts. What, then, is this additional angular momentum?

Let us define a vector $\mathbf{S}$ as follows:

$$S_x=-\frac{\hbar}{2i}\alpha_y\alpha_z, \quad S_y=-\frac{\hbar}{2i}\alpha_z\alpha_x, \quad S_z=-\frac{\hbar}{2i}\alpha_x\alpha_y.$$

Its explicit form can be constructed from the α-matrices. Simple calculation then shows that

$$HS_z-S_zH=-\hbar^2c\left(\alpha_x\frac{\partial}{\partial y}-\alpha_y\frac{\partial}{\partial x}\right).$$

Comparing this with the former result we see that L_z+S_z, and hence the vector $\mathbf{L}+\mathbf{S}$ does commute with H. Therefore S represents what has been called the spin.

The electron possesses also a magnetic moment. To show this requires more detailed calculation, briefly indicated here. If we apply the operator H in Eq. (14) twice, that is, compute $H^2\psi$, the result can be expanded in such a way that $(H^2-\epsilon^2)\psi=0$ represents the Schrödinger equation with additional terms. Among the latter, there appears $(e/2mc)\mathbf{S}\cdot\mathbf{H}$, where $\mathbf{H}=\nabla\times\mathbf{A}$ is the magnetic field strength, a result which implies that the spin $\mathbf{S}$ has associated with it a magnetic moment $(e/2mc)\mathbf{S}$. Hence the theory

explains fully the discovery of Uhlenbeck and Goudsmit.

In concluding the present section, we return to the problem of the free particle, which has some further aspects of great interest. It is to be remembered that in Dirac's theory a simple state is represented by four functions $\psi_1 \cdots \psi_4$. The quadruplet of functions belonging to ϵ as given by Eq. (11) with the positive sign is different from that corresponding to the negative sign. Furthermore, the state for $\epsilon>0$ has a twofold degeneracy, as has also the state for $\epsilon<0$. Thus, to a given value of p there correspond altogether four different states, two belonging to $\epsilon>0$, two to $\epsilon<0$. The first two states can be so combined that the two combinations represent eigenfunctions of S_z, one belonging to the eigenvalue $+\frac{1}{2}\hbar$, the other to the eigenvalue $-\frac{1}{2}\hbar$. A similar linear combination may be effected for the latter two states. We have thus accounted completely for the existing degeneracy; the four states of the electron which are associated with a given linear momentum are to be interpreted as having the properties:

$$\begin{array}{cccc} \epsilon>0 & \epsilon>0 & \epsilon<0 & \epsilon<0 \\ S_z=+\frac{1}{2}\hbar, & S_z=-\frac{1}{2}\hbar, & S_z=+\frac{1}{2}\hbar, & S_z=-\frac{1}{2}\hbar. \end{array}$$

The last two of these are highly paradoxical.

For consider Eq. (12), which is true for all states belonging to a given p. If ϵ is negative, then a *negative* $\bar{v}_x$ corresponds to a positive p. Velocity and momentum would be in opposite directions. If a force (which is here equal to dp/dt, not mdv/dt) causes p to increase, v would decrease in that direction: The electron with negative kinetic energy would be accelerated in a direction opposite to that of the force. Newton's laws of motion would not apply to it. In short, such states are not known to exist.

Dirac has attempted, with partial success, to turn the defect of his theory manifesting itself in this grotesquely false prediction into a virtue by proposing his "hole" hypothesis. This consists essentially of two assumptions:

(1) All states with energies from $-mc^2$ to $-\infty$ are normally occupied.

(2) The electrons filling these states produce no field and make no contribution to the observable aspects of the universe, so long as external fields do not disturb them. All measured electromagnetic quantities refer to the completely filled state as zero level.

Except for assumption (2), the negative-energy electrons would possess not only an infinite negative charge but an infinite charge density at every point of space. Assumption (2) is highly asymmetric, for while it denies negative-energy electrons the possibility to produce external fields, it nevertheless supposes them to be acted on by such fields. Logically, the entire hypothesis is clearly untenable, but curiously it leads to many correct results which, it is generally hoped, will point beyond and possibly correct the theory in the future.

The hypothesis provides an explanation for the existence of positive electrons. For if one of the ubiquitous electrons of energy, say, $-|\epsilon|$, momentum $\mathbf{p}$ and charge $-e$ were missing, the universe would be enriched by an observable energy $+|\epsilon|$, a charge $+e$ and a momentum $-\mathbf{p}$. In other words, a hole in the distribution of negative-energy electrons behaves like a positive electron, with positive energy and momentum opposite to that of the hole. Similarly, if by some external agency a negative-energy electron is raised to a positive energy state, the observable result is the appearance of *two* electrons, one positive and one negative. In this act, the agency must supply the energy difference between the positive and negative states, that is, at least $2mc^2$. Thus the process corresponds in all respects to pair production, which in Dirac's scheme represents itself as a sort of photoelectric effect in which an electron is raised out of the sea of negative energy states. The reverse process corresponds clearly to annihilation of a pair with release of an amount of energy greater than $2mc^2$. Nor does the power of the hole hypothesis exhaust itself in these qualitative explanations; when applied in detail it gives numerically correct values for the probabilities with which the processes occur. It can be shown[12] that pairs can be produced by various agencies, such as gamma-rays in the neighborhood of a nucleus, protons and alpha-particles passing through matter, fast electrons near nuclei, collision of two electrons and collision of light quanta in hot stars. But these matters are beyond the scope of the present report.

[12] See W. Heitler, *Quantum theory of radiation* (Oxford Press, 1936).

5. RADIATION

Atomic and molecular structure are on the whole well accounted for by quantum theory, a theory originally invented by Planck to deal with the problem of radiation. Strangely, its application to this field has ultimately been less successful than to straightforward atom dynamics. This is partly because radiation cannot be completely described without the use of relativistic considerations, which we have seen to be troublesome, and partly because it raises peculiar difficulties of its own.

One may distinguish four states of refinement in the recent development of radiation theory, and these will here be presented in order. The first corresponds to the early formulation of Planck and Bohr with its central recognition of the fact that the frequency of radiation is equal to the energy change suffered by its emittor, divided by Planck's constant h. This law, written in the form

$$E=h\nu, \tag{1}$$

is usually referred to as Bohr's frequency condition; it has occupied our attention before. The mechanism whereby the emitting atom or molecule converted its internal energy into radiation remained entirely obscure; the electron jump was a noncausal, anomalous event which defied detailed explanation. At this level of the theory questions as to frequency only, not concerning the intensity of light, could be answered—except for the one famous case of the blackbody spectrum which, it now appears, the theory could handle by accident.

The transition from this early stage to the next was made possible by the never-failing correspondence principle, whose role was discussed in reference 1. It permitted the calculation of intensities of spectral lines by stipulating agreement with classical electrodynamics in the limit of high quantum numbers, but left largely unsatisfied the curiosity of those desiring fundamental understanding. The first significant advance was made when Schrödinger,[13] after establishing the essential foundations of quantum dynamics, showed how his new scheme could be used to derive, not only the frequencies but also the intensities of spectral lines.

The logic of his procedure is remarkably simple. We illustrate it by considering the problem of the intensity of an absorption line. Monochromatic light of frequency $\nu=\omega/2\pi$, with its electric vector $F_0 \sin \omega t$ along the x axis, progresses along the z axis. It meets an atom whose possible energy states are $E_1, E_2 \cdots$, but which prior to the interaction with the light wave resides in its lowest, or normal, energy state E_1. The intensity of the absorption line is proportional to the energy *lost* by the light wave in a unit of time, and this again is proportional to the probability that the atom has *received* that energy, that is, has transferred itself to a corresponding excited state, say E_k. The problem to be solved may therefore be phrased in essence as follows: What is the probability that, at a time t after the light has been turned on, the atom which originally had an energy E_1 will be found to have a greater total energy E_k?

Let H_a be the Hamiltonian operator for the atom, and let $u_\lambda=\psi_\lambda \exp(-iE_\lambda t/\hbar)$ be the states of the atom which correspond to the energies E_λ. The stationary-state function ψ_λ, which depends only on the space coordinates, satisfies $H_a\psi_\lambda = E_\lambda\psi_\lambda$; u_λ satisfies

$$i\hbar\partial u_\lambda/\partial t = H_a u_\lambda. \tag{2}$$

At the instant $t=0$, when the light was turned on, the state of the atom was certainly u_1. After that, however, it was altered; for its Hamiltonian was no longer H_a, but H_a plus the energy of the light wave. Let us call the new state v. If the contribution of the light to H is denoted by V, then v must satisfy

$$i\hbar\partial v/\partial t = (H_a+V)v. \tag{3}$$

Knowing v, we could expand it in terms of atomic eigenstates u_λ thus:

$$v=\sum_\lambda a_\lambda u_\lambda, \tag{4}$$

where the coefficients a_λ are of course functions of t. Then, according to the axioms of SEC. 1, $|a_k|^2$ represents the probability we are seeking.

So much for the method; let us now fill in some of the mathematics. The energy of the light wave

[13] E. Schrödinger, Ann. d. Physik **81**, 109 (1926).

is easily seen to be[14]

$$V=-eF_0x\sin\omega t. \tag{5}$$

This energy is assumed to be so small that V^2 can always be neglected in comparison with V. In order to solve Eq. (3) it is well to expand v in accordance with Eq. (4) at once. On simple calculation, this change converts Eq. (3) into a set of linear differential equations in t, involving all the coefficients a_λ:

$$\dot{a}_k=-\frac{i}{\hbar}\sum_\lambda a_\lambda V_{k\lambda}\left[\exp\left(i\frac{E_k-E_\lambda}{\hbar}t\right)\right], \qquad k=1,\ 2,\ \cdots, \tag{6}$$

where

$$V_{k\lambda}\equiv\int\psi_k^*V\psi_\lambda d\tau.$$

Since at $t=0$ all coefficients a_λ with the exception of a_1 were zero, and a_1 was 1 at that instant, we proceed to assume these values for all times here considered in evaluating the sum over λ in Eq. (6), which therefore reduces to a single term. The result thus obtained is valid only for small values of t. On writing now ω_k for $(E_k-E_1)/\hbar$, this quantity being indeed the frequency which the atom would emit in the transition from E_k to E_1 by virtue of Eq. (1), Eq. (6) becomes

$$\dot{a}_k=\frac{eF_0}{2\hbar}x_{k1}\left[e^{i(\omega_k+\omega)t}-e^{i(\omega_k-\omega)t}\right]$$

when $\sin\omega t$ in Eq. (5) has been expanded; x_{k1} is Heisenberg's coordinate matrix element associated with the (hypothetical) transition from E_k to E_1. When this expression for $\dot{a}_k$ is integrated and the constant of integration so determined that $a_k=0$ when $t=0$, we find

$$a_k=\frac{ieF_0}{2\hbar}x_{k1}\frac{e^{i(\omega_k-\omega)t}-1}{\omega_k-\omega}+\cdots,$$

where a similar term with a denominator $\omega_k+\omega$ has been neglected. The reason for this drastic curtailment is that in the present connection we are interested only in frequencies ω which lie in the neighborhood of ω_k, and for all these the omitted terms are insignificant. The final result then takes the form

$$|a_k|^2=\frac{1}{2\hbar^2}\cdot F_0^2\cdot|x_{k1}|^2\cdot\frac{1-\cos(\omega_k-\omega)t}{(\omega_k-\omega)^2}. \tag{7}$$

This result is so illuminating, so laden with information, that we have not hesitated to insert the few steps leading to it even in this rapid survey. It marks the second stage of refinement through which radiation theory has evolved. As the reader will recall, $|a_k|^2$ represents the probability of absorption in a *short* time t.

Aside from the constant factor $(2\hbar^2)^{-1}$, $|a_k|^2$ depends:

(1) On F_0^2, implying that, as in classical physics, absorption intensity is proportional to the square of the electric vector, that is, to the intensity of the incident wave. This is nothing new.

(2) On $|x_{k1}|^2$. Here we have in a nutshell both the "selection rule" and the "polarization rules" of the older quantum theory, which, being unable to prove them, *postulated* both. For this statement means that the probability of absorption is zero with respect to all atomic transitions $E_1\rightarrow E_k$ whose matrix element x_{k1} vanishes, which is the old selection rule. Had we assumed the incident light to be polarized along y or z, $|x_{k1}|^2$ in Eq. (7) would be replaced by $|y_{k1}|^2$ or $|z_{k1}|^2$. Hence only that component of polarization will be absorbed for which the corresponding coordinate matrix element is finite. This, when phrased for emission of light rather than absorption, is the usual polarization rule.

(3) On $[1-\cos(\omega_k-\omega)t](\omega_k-\omega)^{-2}$, a factor showing an interesting dependence on the frequency of the incident light. Aside from other minor maximums at an infinite number of places, it has a sharp peak when $\omega=\omega_k$, and this peak grows with increasing t. If t is sufficiently large it is the only one of importance. But the condition $\omega=\omega_k$ is simply the Bohr frequency condition, Eq. (1), which is thus automatically accounted for by the theory. Alternatively, if we assume

[14] Such a formula can of course not be justified rigorously. The light wave has no scalar potential of the form here written. However, if the somewhat more complicated expression involving the vector potential were written, the final result would be the same. Equation (5) really represents a mechanical force urging the atomic charge back and forth with the frequency of the light wave. For our present purpose this fiction is acceptable.

Planck's law, $E=h\nu$, the condition appears as the statement of conservation of energy.

We may not, however, make t as large as we please because, as was noted, the whole approximation is limited to small values of t. Thus the peak is not infinitely sharp. Does this mean that energy is not exactly conserved? We can hardly draw such an inference; it is the uncertainty principle that obtrudes itself into the situation at this point. For t and E are canonically conjugate variables, so that $\Delta t \cdot \Delta E \approx \hbar$. In the problem under consideration, t must be regarded as Δt, the interval of waiting within which the energy is uncertain, so that $\Delta E \approx \Delta t/\hbar$. If t is not large, we do not know the value of the energy precisely, and this is why its conservation cannot be stated with precision.

Thus far the incident light has been assumed to be monochromatic. In general, it will have a *range* of frequencies $\Delta\omega$ over which the intensity is a slowly varying function $f(\omega)$. The probability given by Eq. (7) must then be integrated over $d\omega$ after multiplication by $f(\omega)$. Now

$$\int_{\Delta\omega} f(\omega)\frac{1-\cos(\omega_k-\omega)t}{(\omega_k-\omega)^2}d\omega = f(\omega_k)\cdot\pi t.$$

The probability in question, for small t, is indeed proportional to t and we can compute the probability per unit time:

$$w=\frac{\pi F_0^2}{2\hbar}|x_{k1}|^2 f(\omega_k).$$

When the Planck distribution function, which describes the density of radiation in equilibrium with matter, is inserted for $f(\omega)$ in this formula, there results a well-substantiated expression for the intensity of a spectral absorption line.

The scheme developed here lends itself readily to the treatment of other problems, notably those of refraction and dispersion, where it meets with complete success.

But its major stumbling block is the incomplete account it is forced to give of the phenomena of *emission* of light. If the atom is initially in an excited state u_k and there is no light wave present, so that $V=0$, Eq. (2) continues to be satisfied for all times; the state is a stationary one and no light will ever be emitted, which is of course quite contrary to the facts. The error can be corrected in an artificial way by introducing an extraneous piece of knowledge gleaned from thermodynamics, namely, the relation between the probabilities of emission and of absorption (Einstein's well-known coefficients A and B). It is this need of augmentation, the theory's failure to explain emission in the same manner as absorption, which calls for thoroughgoing revision and takes us to the third stage of refinement.

This third stage was developed by a considerable number of authors, chiefly by Dirac,[15] Heisenberg and Pauli,[16] and Fermi.[17] They saw the roots of the defect of the foregoing analysis in the circumstance that it treats light *only* as a perturbation on matter. Obviously, a space filled with radiation—with photons—is a physical system just as definite and important as an assemblage of atoms or molecules. But the former theory has no representation for it. Hence a way must first be found to describe a radiation field in quantum fashion.

Presumably, the radiation field has associated with it a Hamiltonian operator H_r. An atom embedded in a radiation field will then be a composite system corresponding to a Hamiltonian operator

$$H=H_a+H_r+V, \tag{8}$$

where V is the interaction energy between atom and radiation. The equation

$$(H_a+H_r)u=i\hbar\partial u/\partial t \tag{9}$$

can be solved, as we shall see. The "perturbed" equation

$$Hv=i\hbar\partial v/\partial t \tag{10}$$

can then be treated by the same approximation scheme as before. But let us first construct H_r.

Limitation of space will force us to omit some steps. In a pure radiation field, the scalar potential $\varphi=0$ and the vector potential $\mathbf{A}$ satisfies the wave equation

$$\left.\begin{aligned}\nabla^2\mathbf{A}-\frac{1}{c^2}\frac{\partial^2\mathbf{A}}{\partial t^2}=0,\\ \text{also}\qquad \nabla\cdot\mathbf{A}=0.\end{aligned}\right\} \tag{11}$$

[15] P. A. M. Dirac, Proc. Roy. Soc. **114**, 243 (1927).
[16] W. Heisenberg and W. Pauli, Zeits. f. Physik **56**, 1 (1929); **59**, 169 (1930).
[17] E. Fermi, Rev. Mod. Phys. **4**, 131 (1932).

If the radiation field exists in a space of finite volume, on the surface of which $\mathbf{A}$ must vanish, then $\mathbf{A}$ may be developed as a series of *orthogonal* functions $\mathbf{A}_\lambda$ depending only on space coordinates and satisfying

$$\nabla^2\mathbf{A}_\lambda+\frac{\omega_\lambda^2}{c^2}\mathbf{A}_\lambda=0 \tag{12}$$

in this way:

$$\mathbf{A}=\sum_\lambda q_\lambda\mathbf{A}_\lambda. \tag{13}$$

The coefficients q_λ are functions only of the time. When Eqs. (13) and (12) are substituted into the first of Eqs. (11), every q_λ is seen to satisfy the equation

$$\frac{d^2q_\lambda}{dt^2}+\omega_\lambda^2q_\lambda=0, \tag{14}$$

which represents a simple harmonic oscillator vibrating with angular frequency ω_λ.

Leaving the mathematics aside, we see that the present result permits the following interpretation. Note first that the quantities ω_λ and $\mathbf{A}_\lambda$ of Eq. (12) are not, in a sense, descriptive of the radiation field under study; they are determined by the enclosure and represent predetermined forms into which the properties of the field must fit. The contingent characteristics of the radiation actually present are entirely given by the q_λ whose temporal behavior is dictated by Eq. (14), but whose actual magnitudes make the field what it is. For example, many of them may be zero. It is therefore quite proper for us to regard the field as essentially represented, even in *classical* electrodynamics, by an infinite set of simple harmonic oscillators q_λ, whose amplitudes can be adjusted to suit all possible conditions. It is this fact that forms the link between classical and quantum radiation theory.

The *energy* of the radiation field, when written in terms of electric and magnetic field strengths $\boldsymbol{E}$ and $\boldsymbol{H}$, is

$$\frac{1}{8\pi}\int(\boldsymbol{E}^2+\boldsymbol{H}^2)d\tau.$$

When $\boldsymbol{E}$ and $\boldsymbol{H}$ in this formula are replaced by their well-known expressions in terms of $\mathbf{A}$, and $\mathbf{A}$ in terms of the q_λ by means of Eq. (13), the volume integral turns into

$$\tfrac{1}{2}\sum_\lambda\left[\left(\frac{dq_\lambda}{dt}\right)^2+\omega_\lambda^2q_\lambda^2\right], \tag{15}$$

provided use is made of the fact, not previously stated, that

$$\int\mathbf{A}_\lambda\cdot\mathbf{A}_\mu d\tau=4\pi c^2\delta_{\lambda\mu}.$$

Now expression (15) represents simply the energy of a set of linear simple harmonic oscillators; it shows that our interpretation still holds. But it also indicates how the transition to quantum mechanics is to be made. We invoke nothing but the axioms outlined in SEC. 1. The result (15) may be written in Hamiltonian form,

$$H_r=\tfrac{1}{2}\sum_\lambda(p_\lambda^2+\omega_\lambda^2q_\lambda^2)\equiv\sum_\lambda H_\lambda, \tag{16}$$

as an application of Hamilton's canonical equations will immediately prove. If we regard the q_λ as coordinates and replace p_λ by $-i\hbar\partial/\partial q_\lambda$, the equation

$$H_\lambda\psi_\lambda=E_\lambda\psi_\lambda$$

is the familiar Schrödinger equation for a simple harmonic oscillator of unit mass. Its solutions are Hermite functions; the eigenvalues[18] are given by

$$E_\lambda=(n_\lambda+\tfrac{1}{2})\hbar\omega_\lambda,$$

n_λ being an integer. The energy of the whole field is clearly

$$\sum_\lambda E_\lambda=\sum_\lambda(n_\lambda+\tfrac{1}{2})\hbar\omega_\lambda. \tag{17}$$

Radiation thus appears as an infinite system of oscillators (whose locations are of course not specified since the "space" of the q_λ is not ordinary space!), each vibrating with its own characteristic frequency ω_λ. Their energies are quantized, and n_λ is the number of quanta possessed by the λth oscillator. If all n_λ are zero, the field is dark.

One would expect the energy of the dark space to be zero. Equation (17), however, shows it to be

$$\sum_\lambda\tfrac{1}{2}(\hbar\omega_\lambda)=\infty.$$

The radiation field has an infinite "self-energy." This peculiar and unwelcome proposition has no physical significance. Its emergence is nevertheless interesting because it may be related to other difficulties to which the present theory gives rise.

[18] See SEC. 1, Eq. (5).

The infinite self-energy can be avoided by making the transition from classical to quantum mechanics in a slightly different way, but the theory has no self-regulating mechanism that forces its avoidance.

At this stage of the theory, the p_λ's and q_λ's no longer commute. Since $\boldsymbol{E}$ and $\boldsymbol{H}$ are functions of these operators, they, too, will no longer commute in general. The investigation of their behavior is the subject of quantum electrodynamics, to which the papers by Heisenberg and Pauli (reference 16) are largely devoted.

Having quantized the radiation field, that is, having found the operator H_r [see Eq. (16)] which occurred in Eq. (8), we are ready to turn briefly to the general problem of finding out what radiation does to atoms. Here we can be brief, for the logic of this problem is simple though its solution is cumbersome. The solution of Eq. (9) is now available; it is the product of the solution u_a of the Schrödinger equation for the atom,

$$H_a u_a = i\hbar \partial u_a/\partial t,$$

and the solutions u_λ for the radiation oscillators, satisfying

$$H_\lambda u_\lambda = i\hbar \partial u/\partial t.$$

In other words,

$$u = u_a \prod_\lambda u_\lambda.$$

Here it is to be recalled that u_a is a function of $\mathbf{r}$, the atomic (electron!) coordinates, and each u_λ is a function of q_λ, the (generalized) coordinate representing the λth oscillator.

But what is V, the interaction energy occurring in Eq. (8)? In classical theory it has the form $-(e/m)\mathbf{p}\cdot\mathbf{A}$, as follows from the developments in the preceding section. Here $\mathbf{p}$ is the momentum operator of the atomic electron. Therefore V depends on both the variables contained in u_a and those in all u_λ, the latter because $\mathbf{A}$ is a function of all q_λ. We have thus arrived at the point where the equation

$$(H_a + H_r + V)v = i\hbar \partial v/\partial t$$

may be solved by the method applied to Eq. (3), $H_a + H_r$ now taking the place of the former H_a. To be sure, V now has a different form, depending in fact no longer on t. But the mathematics is otherwise the same, involving in particular the use of Eq. (6). The result for the absorption problem is identical with that derived by the simpler method, and the method describes emission correctly also.

And it achieves much more than that. In one of its earliest applications Weisskoff and Wigner[19] were able to account for the natural width of spectral lines. Dispersion, Raman effect, photoelectric emission, Compton effect and many nuclear phenomena involving gamma-rays were correctly explained in quantitative detail. The successes of the theory at this stage are most impressive. For further discussion we must refer the reader to special treatises, such as Heitler's book.[12]

Nevertheless, the radiation theory contains grave paradoxes which will here be mentioned in passing. Some of them may be traced to relativistic sources and are mere illustrations of the difficulties alluded to in Sec. 4. They result in the theory's failure to be applicable to wave-lengths smaller than $\hbar/mc$; the hole hypothesis causes trouble since it converts a vacuum into a sea of electrons capable of being acted on by radiation. In addition to these, radiation theory creates one major difficulty of its own. We have seen that, in the scheme here used, the energy of the radiation field is infinite even if no radiation is present. It was pointed out, however, that by the use of a mathematical trick this can be avoided; hence the "zero-point" energy of a pure radiation field can be eliminated. But this is not true when the atom is present! For the interaction term V makes a contribution to the energy that cannot be ignored. If this contribution is computed in accordance with certain well-known formulas of perturbation theory, the first perturbation is

$$V^{(1)} = \bar{V} = 0,$$

which is satisfactory. But the second perturbation,

$$V^{(2)} = \sum_k \frac{|V_{0k}|^2}{E_0 - E_k},$$

diverges, even when the radiation field is dark. Calculation shows this divergence to be connected with the possible presence in the field of

[19] V. Weisskoff and E. Wigner, Zeits. f. Physik **63**, 54 (1930).

photons having very high energy. For this reason, theories have been proposed which legislate such photons out of existence. But from a logical point of view, these attempts perhaps do not warrant further discussion.

As to the fourth stage of radiation theory, it has not produced significant results other than those already outlined. It is a somewhat more elegant method, known as *second quantization*, of deriving them. Since it contains methodological elements of a rather different sort, we shall outline it very briefly.[20]

Let us return to Eq. (6), which we now write in the form

$$i\hbar\partial\alpha_k/\partial t=\sum_\lambda H_{k\lambda}\alpha_\lambda. \tag{18}$$

Formerly, a_k was the amplitude with which the kth atomic state was present in the perturbed state function of the atom. Now α_k represents the amplitude of the kth state of the whole radiation field. The index k therefore represents the manifold of oscillator quantum numbers $n_1, n_2, \cdots$, and α_k may therefore be written as a *function* of these quantum numbers: $\alpha(n_1, n_2, n_3, \cdots)$. Furthermore, $H_{k\lambda}=\int\psi_k^*H\psi_\lambda d\tau$, where ψ_k and ψ_λ describe the entire radiation field, and H is the total Hamiltonian operator, containing the oscillator Hamiltonians of all the photons. We now introduce the operator $H^{(1)}$ which refers to a *single* photon, a much simpler expression. By a tedious analysis it may be shown that Eq. (18) is entirely equivalent to

$$i\hbar\frac{d}{dt}\alpha(n_1\cdots n_r\cdots)=\sum_{st}H_{st}^{(1)}b_sb_t^\dagger\alpha(n_1\cdots n_r\cdots), \tag{19}$$

where $H_{st}^{(1)}=\int\psi_s^*H^{(1)}\psi_t d\tau$ with ψ_s and ψ_t taken as state functions of a *single* photon, provided b_s and $b_t^\dagger$ are *operators* acting on the "variables" $n_1, n_2, \cdots$ appearing in α. These operators must satisfy the commutation rules

$$b_sb_t-b_tb_s=0,$$
$$b_s^\dagger b_t-b_tb_s^\dagger=\delta_{st};$$

they can be easily represented as matrices. Equation (19), although it appears more complicated, is actually easier to handle than Eq. (18). But we can go one step further. If we introduce a new state function

$$\Psi(\mathbf{r})=\sum_s b_s^\dagger\psi_s^*(\mathbf{r})$$

whose argument $\mathbf{r}$ has reference to the coordinates of a single photon, and which is constructed through an ordinary expansion in terms of the orthonormal ψ_s, but with *operators* $b_s^\dagger$ instead of numbers as usual, then Eq. (19) takes the form

$$H^{(1)}\Psi(\mathbf{r})\alpha(n_1\cdots n_r\cdots)=E\Psi(\mathbf{r})\alpha(n_1\cdots n_r\cdots),$$

which is interesting indeed. For instead of the normal equation,

$$H\Psi=E\Psi,$$

where H acts on the coordinates of Ψ, we have here an operator $H^{(1)}$ which affects the coordinate part of $\Psi(\mathbf{r})$, and a second operator $\Psi(\mathbf{r})$ (by virtue of the b's) which acts on the arguments of α. This circumstance has given the method its name—second quantization. While some authors believe it to be the vehicle by which radiation theory will be advanced in the future, others feel that its usefulness is limited.

[*This is the second of three articles on atomic and molecular theory since Bohr.*]

[20] See P. Jordan and E. Wigner, Zeits. f. Physik **47**, 631 (1928); V. Fock, Zeits. f. Physik **75**, 622 (1932).

Revised List of Available Reprints

REPRINTS of the following articles and reports are still available and may be purchased at cost from the Editor, AMERICAN JOURNAL OF PHYSICS, Wabash College, Crawfordsville, Indiana. Orders will be accepted only from subscribers and, in the case of articles, only for six or more copies. Payment may be made in stamps.

Margenau and Wightman, *Atomic and molecular theory since Bohr: historical survey*, 35 cts for 6 copies.

Behren, *Atomic theory from 1904 to 1913*, 25 cts for 6 copies.
Behren, *Early development of the Bohr atom*, 40 cts for 6 copies.
Behren, *Further development of Bohr's early atomic theories*, 30 cts for 6 copies.
Weisskopf, *On the theory of the electric resistance of metals*, 40 cts for 6 copies.
Hamilton, *Molecular beams and nuclear moments*, 60 cts for 6 copies.
Kerst, *The betatron*, 35 cts for 6 copies.
AAPT Committee, *Proposal to standardize letter symbols*, 10 cts per copy.
APS Committee, *Physics in relation to medicine*, 10 cts per copy.

Ballistics of Small-Arms Ammunition

RALPH HOYT BACON AND WILLIAM J. KROEGER
Frankford Arsenal, Philadelphia, Pennsylvania

THE subject of ballistics is traditionally divided into two parts—interior ballistics and exterior ballistics. *Interior ballistics* is the study of the phenomena from the start of the ignition of the primer until the escape of the projectile from the muzzle of the gun; it is a complicated subject, involving chemical kinetics and thermodynamics. The motions of the projectile, gun and powder gases, as well as the flash and smoke at the muzzle of the gun, are all included in the field of interior ballistics. The subject is divided into two main parts: primer phenomena; and propellant powder phenomena, or interior ballistics proper, in the older sense of the word.

Exterior ballistics is the study of the flight of the projectile from the muzzle to the target. It is a highly specialized field of analytic mechanics.

Today, there is added a third section, called *terminal ballistics*, or the effect of the projectile on the target.

A typical small-arms cartridge is shown in Fig. 1. Its chief components are: (i) the *case* (this is the usual military term; popularly, it is called the shell); (ii) the *primer;* (iii) the *propellent powder;* (iv) the *bullet.*

The cavity into which the primer fits is called the *primer pocket;* the hole between the pocket and the interior of the case is the *vent;* the end of the case containing the primer is the *head;* the end containing the bullet is the *mouth.* These are the military terms. In sporting circles, other names are sometimes used.

THE PRIMER

In U. S. Army small-arms ammunition, the primer generally consists of three components: (i) the *primer cup;* (ii) the *anvil;* (iii) the *priming mixture.* A typical small-arms primer is shown in Fig. 1.

The primer cup is made of soft metal, such as brass or gilding metal. The anvil is made of harder brass. The impact of the firing pin of the gun against the cup produces an indent, and this, in turn, transmits much of the energy of the incident firing pin to the very small portion of the priming mixture between this indent and the point of the anvil, raising the temperature of this bit of priming mixture to the ignition point. The priming mixture, in many military primers, is a mixture of potassium chlorate and TNT. The flame produced in the small portion of the priming mixture between the anvil point and the indent spreads throughout the rest of the priming mixture with incredible violence and speed, and ultimately (in a few microseconds) leaps through the vent deep into the propellent charge inside the case.

The two most important properties of primers are sensitivity and uniformity of ignition.

The Primer Sensitivity Test

The sensitivity of primers must be kept under close control. If primers become too sensitive, they are dangerous to handle during manufacture and packing of the cartridges; if they are not sensitive enough, there will result a probability of misfires in the guns.

The sensitivity of primers is measured as follows. A steel ball is allowed to fall through a known height onto the head of a pin, the contour of whose point is that of the firing pin of a weapon. The impact of this ball causes the pin to strike the primer in much the same way that the firing pin of a weapon would do it. Generally, 50 primers are tested at each height of fall. At the lower intermediate heights, some of the primers will fire and others will not; and, of each sample tested, the fraction firing will be denoted by P. This test is repeated for a few heights of fall, and the fraction firing at each height of drop is recorded.

Generally, the primers to be tested are inserted into cartridge cases, in exactly the same way as for service ammunition, since it is found that the measured sensitivity is a function of the manner of insertion, and of the firmness and rigidity with which the case is held in the primer drop apparatus.

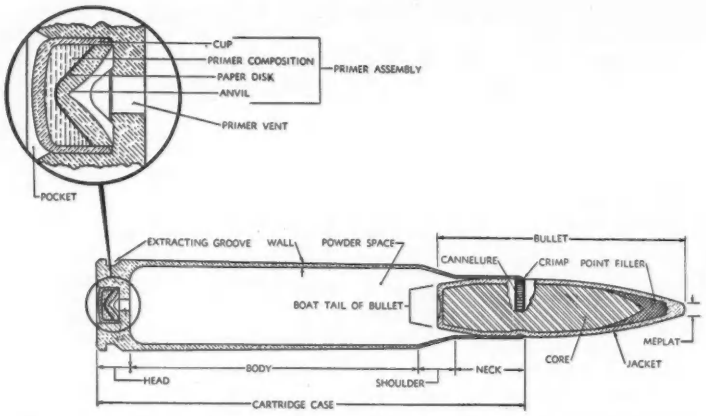


FIG. 1. Typical military small-arms cartridge. The cartridge shown contains an armor-piercing bullet. The usual small-arms primer is shown in the insert. The paper disk is to prevent the primer composition from extruding out between the legs of the anvil, which is inserted into the primer while the mixture is still soft and plastic.

Let H be the height of drop, and let $f(H)dH$ be the number of primers requiring a height between H and $H+dH$ in order to be fired. It is impossible to find the exact height of drop required to fire any individual primer, but the fraction P firing at any given height is

$$\int_0^H f(H)dH.$$

For most practical purposes, the fraction P firing at each height H may be considered to lie on the ogive of the normal error curve; that is, for a first approximation, we have

$$P=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{H} e^{-(H-\bar{H})^2/2\sigma^2}dH.$$

The sensitivity of each lot of primers may therefore be described by two parameters: $\bar{H}$, the height at which 50 percent will fire, and σ, the standard deviation of the sensitivity. The energy of impact of the firing pin of a gun corresponds to such a large value of H that the probability of a misfire is only one out of several million.

Actually, of course, the points representing P and H cannot lie exactly on the ogive of the normal error curve, since this curve passes above the origin of coordinates, whereas the true sensitivity curve must either pass through the origin, or, what is more likely, pass below it, though the portion of the curve below the H-axis would not have physical significance. It is found that the first two terms of the Gram-Charlier series, or a Pearson Type-III curve, represent the sensitivity sufficiently close for even the most critical work. This means that a third parameter, called the *skewness* and denoted by α_3, is involved in the complete description of the sensitivity of a lot of primers. However, the ordinary test sample is not usually large enough to determine α_3 accurately, and therefore, only $\bar{H}$ and σ are usually determined.

Although the primer drop test is very old, the analysis and interpretation given here are recent, and are due to Dr. C. W. Churchman of this arsenal.

The Ignition Test

One of the tests commonly employed to determine the uniformity of ignition of primers is called the *hangfire test*. The current apparatus for making this test, called the *electrostatic hangfire recorder*, depends on the fact, first noted at this arsenal, that when the bullet and powder gases leave the muzzle of the gun, they are electrostatically charged. An insulated metal ring is mounted just in front of the gun and connected to a suitable amplifier. The passage of the bullet and gases through the plane of this ring sends an impulse, or signal, through the amplifier, and

this amplified signal is applied to the grid of a thyratron, rendering the tube conducting. This, in turn, discharges a condenser through the primary of a spark coil, thus causing a spark to jump to a rotating drum covered with stylograph paper and driven by a synchronous motor, which also controls the firing of the gun.

When a hangfire test is being conducted, a spark to mark the initiation of the firing cycle is first caused to jump to the rotating drum. Then follow the sparks caused by the passage of the bullet and gases through the insulated ring. We can thus measure the total time required for the fall of the firing pin, the ignition of the primer and of the propellent powder, and the travel of the bullet through the bore of the gun. This is done for a few hundred cartridges of each lot.

The actual length of these firing cycles will depend upon the characteristics of the gun as well as upon those of the cartridge. An excessively long firing cycle—for example, one more than 1.4 msec longer than the shortest cycle—is called a *hangfire*. A hangfire is usually, though not always, attributable to a faulty primer. Hangfires are not allowed in military ammunition.

INTERIOR BALLISTICS

The propellent charge in military small-arms ammunition is usually a mixture of cellulose nitrates together with other chemicals whose purpose is to regulate the rate of burning of the charge, to lower the temperature of the products of combustion, to reduce the flash at the muzzle, to increase the storage life (chemical stability) and to decrease the hygroscopicity. Approximately 1 percent of the weight of the charge is residual moisture left over from the manufacturing process, and it is desirable that this moisture remain reasonably constant in order that uniform pressures and velocities may be obtained.

The size and shape of the powder grain, the proportions of moisture and other chemicals, and so on, are all very critical. The powder must burn at the right speed, or else either the pressure will be undesirably high or the velocity imparted to the bullet will be too low. The temperature of the products of combustion must not be too high, or else the rate of erosion of the gun will be too great. Therefore, for best results, each different type of cartridge requires a special powder. However, it is possible, and it is the practice, to load reasonably similar cartridges with one type of powder.

For small-arms powders, the easiest and most satisfactory way of controlling the rate of burning and the resultant temperature and pressure is by the addition of certain coating agents after the powder grains have been cut to size. It is possible to make powders suitable for widely different purposes from the same base grain by carefully controlling the coating process. However, excessive coating must be avoided, as powders with heavy coatings do not ignite uniformly, and the dispersions in the velocities produced are undesirably large. Heavily coated powders are oversensitive to variations in the dimensions of rifle barrels and of bullets.

The Pressure Gage

The pressure generated by the burning propellent charge in a small-arms cartridge is usually measured in a modified weapon such as is shown in Fig. 2. A yoke is placed over the barrel, and a hole is drilled through the yoke and barrel into the side wall of the chamber. Generally, the area of this hole is 1/30 in.^{2}. A steel piston is then fitted to this hole. A soft annealed copper cylinder is placed on the head of the piston, and held there by the screw from the top of the yoke.

For routine manufacturing tests, a small copper cup at the foot of the piston acts as a valve or gas check. For measuring the pressures of certain cartridges, a hole is drilled in the side of the cartridge case before inserting the cartridge into the chamber of the gun. Many other types of cartridge are inserted into the chamber intact, and when the round is fired, a small piece of metal is torn from the cartridge case and blown into the piston hole against the gas check by the burning gases. For more important or critical work, a Neoprene plug or disk, instead of the copper cup, is used as a gas check, and the hole in the cartridge case must be very carefully alined with the piston hole.

The pressure of the burning powder gases forces the piston upward, and this compresses

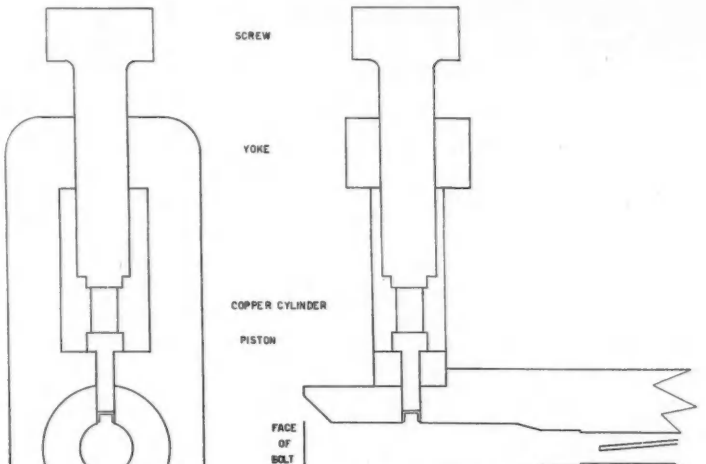


FIG. 2. Section of the chamber of a weapon modified to measure the pressure developed when firing caliber .30 ammunition. The distance from the face of the bolt to the shoulder is called the *breeching space*, or, frequently, the *head space*. This must be accurately adjusted to the right amount when assembling the barrel to the action or receiver. (When a barrel is worn out, it is unscrewed from the action or receiver, and a new one put in its place.) The other dimensions of the chamber cannot be adjusted at assembly; they must be correct beforehand.

the copper cylinder to some fraction (generally about $\frac{7}{8}$) of its original length, depending on the type of cartridge and on the type of copper cylinder. For many types of cartridge, the pressure is between 3000 and 4000 atm.

The quantity measured by the shortening of the copper cylinder is not, of course, the true pressure of the burning powder gases, but rather, a complicated function of the maximum pressure and of the time integral of this pressure. Although the term "pressure" is applied by ammunition manufacturers to this quantity, it is merely a number which informs one, with good assurance, whether the true pressure is dangerous or not.

To measure the true pressure, one merely has to substitute for the copper cylinder a properly calibrated quartz piezoelectric crystal, connected through a suitable amplifier to a cathode-ray oscilloscope. With such an apparatus, one not only can measure the maximum pressure produced, but can also determine the shape of the pressure-time curve.

The Velocity-Pressure Relation
The Efficiency of the Burning Powder

One might think that a given powder charge would always impart the same velocity to a given bullet, or the same kinetic energy to bullets of different masses. This is far from true. The efficiency with which the powder burns depends upon the pressure developed.

As a striking example, consider two service bullets—one, the 150-grain ball caliber .30 bullet well known to hunters and marksmen; the other, the 160-grain armor-piercing bullet. Like charges of the current rifle powders will give practically identical velocities to the two bullets, but the pressure developed by the burning powder gases will be much higher in the second case than in the first (actually, the ratio of the pressures is much larger than the ratio of the masses of the bullets).

In the ball cartridge, there is about $\frac{1}{4}$ in. between the base of the bullet and the powder charge below when the cartridge is stood on the head of the case. If such a cartridge is fired

with the powder packed against the primer, the resulting velocity and pressure will be higher than if the powder is as far from the primer as possible, and packed against the base of the bullet.

The space between the powder and the bullet is called the *air space*. About 0.1 in. air space is needed to allow for variations in the density of the powder and in the available volume of the cartridge cases when loading ammunition by mass-production methods. The excessive air space in the ball cartridge results from the desire to load as many types of cartridge as possible with one type of powder, and this cartridge is loaded with the powder designed for the longer and heavier armor-piercing bullet, which extends further into the cartridge case than does the ball bullet. Part of the increase of pressure observed in the armor-piercing cartridge is the result of the greater length of the armor-piercing bullet and the reduced initial volume available to the burning powder gases.

The variations in the available volumes of the cartridge cases cause corresponding variations in velocity and pressure, higher velocities and pressures being obtained with cartridges the internal volume of whose cases is smaller. The variations in volume are caused by small variations in length, diameter and wall thickness of the cartridge cases, and in the length and depth of insertion of the bullets. Such variations are inevitable in mass production of anything. Similar variations also occur in the diameters of the bullets. For these reasons, components to be used for assembling cartridges for testing propellent powders must be selected with care.

Similarly, the velocity and the pressure will depend upon the primer. Primers of one type cause the powder to burn faster than do those of another type. This is believed to be due, at least in part, to the difference in the violence with which the primer flame leaps through the vent. It is believed that the flame of certain primers actually shatters some of the powder grains, causing them to burn faster than the remaining unbroken ones; this, in turn, affects the velocity and pressure obtained with a given charge and a given bullet. However, the change from one primer to another will not always affect the velocity and pressure alike. Thus, the optimum primer for one cartridge may not be suitable for another.

The variation of velocity and pressure with the position of the powder is believed to be related to the brisance (or violence) of the primer, more powder grains being shattered when they are packed against the primer end of the case than when they are crowded against the base of the bullet. The effects are usually dependent on the temperature of the powder at the instant of firing the primer, and, curiously enough, excessive pressures are occasionally encountered when ammunition is fired at very low temperatures.

The chambers of no two rifles will be exactly alike: There will be small differences of the order of 10^{-3} or 10^{-4} in. in the dimensions of the chamber, and in those of the forcing cone, or region in which the lands of the rifling rise up to their full height. In general, higher pressures and velocities will be obtained in the rifle having the smaller chamber and the shorter forcing cone. For this reason, weapons to be used for testing and standardizing ammunition and propellent powders must be selected and measured with the greatest care.

EXTERIOR BALLISTICS

Let $d^2\mathbf{r}/dt^2$ be the acceleration of a body as viewed in a Galilean coordinate system, and let $\delta\mathbf{r}/\delta t^2$ be the acceleration of the same body as viewed in a coordinate system having the same origin as the Galilean system but rotating with angular velocity ω. Then it is shown in any book on vector analysis that

$$\frac{d^2\mathbf{r}}{dt^2}=\frac{\delta^2\mathbf{r}}{\delta.^2}+2\omega\times\frac{\delta\mathbf{r}}{\delta t}+\omega\times(\omega\times\mathbf{r})+\frac{\delta\omega}{\delta t}\times\mathbf{r}.$$

In the present instance, the rotating coordinate system is the earth, whose angular velocity is constant. The last term in the foregoing equation therefore vanishes. Also, since there are 86,164 sec in a sidereal day, $\omega=2\pi/86{,}164=7.3\times10^{-5}$ sec^{-1} and $\omega^2=5.3\times10^{-9}$ sec^{-2}; hence the terms involving ω and ω^2 may be omitted, except in the case of long range artillery fire. We therefore have, for small-arms ammunition,

$$\frac{d^2\mathbf{r}}{dt^2}=\frac{\delta^2\mathbf{r}}{\delta t^2}.$$

We may therefore describe the motion of the bullet with reference to a right-handed Cartesian

coordinate system with origin at the muzzle of the gun, in which the x-axis is in the direction of the horizontal component of the initial motion of the projectile, the y-axis is the vertical line through the muzzle, and the z-axis is normal to the xy plane, which contains the gun.

When the bullet leaves the gun, it has a large linear velocity (up to 3000 ft/sec) and a tremendous spin (up to 3600 rev/sec, or 216,000 rev/min). The axis of this spin will not, in general, exactly coincide with the tangent to the trajectory. The angle between the axis of spin and the tangent to the trajectory is called the *yaw*, and is denoted by δ. When δ is sufficiently large, the axis of spin will precess about the tangent to the trajectory, and the bullet will also nutate, until δ is reduced by the damping effect of the air friction to a steady value determined by the curvature of the trajectory and by the rate of spin.

Since the axis of spin is not tangent to the trajectory even in the steady state, the total wind resistance **R** may be resolved into two components: the drag **D** tangent to the trajectory, opposite in direction to the velocity of the bullet; and the cross-wind force **L** perpendicular to **D**. Thus the equations of motion become

$$m\mathrm{d}^2x/\mathrm{d}t^2 = -D_x+L_x = -R_x,$$
$$m\mathrm{d}^2y/\mathrm{d}t^2 = -D_y+L_y-mg = -R_y-W,$$
$$m\mathrm{d}^2z/\mathrm{d}t^2 = -D_z+L_z = -R_z,$$

where $W[=mg]$ is the weight of the bullet. From these equations it is seen that the path of a precessing bullet is a helix of some kind; after the precession is damped out, the path is a curve which may, for most purposes, be considered to lie in a vertical plane (generally the xy-plane), since **L** is small; but for some purposes, such as the accurate setting of gun sights, the z-components of the motion cannot be neglected.

The various components of **R** will, in general, be different functions of the velocity through the air, of the air density and of the angle of yaw δ. They will also be functions of the mass and shape of the bullet.

The wind resistance **R** and the weight W of the bullet will not, in general, act through the same point. They will therefore produce a torque called the *overturning moment* **M**, equal to $\mathbf{q}\times\mathbf{R}$, where **q** is the distance between the center of mass and the center of pressure. Since **R** is a function of the velocity and of δ, this expression for **M** is generally written in the scalar form $M=\mu \sin\delta$, where μ is called the *moment factor*. If A denotes the longitudinal moment of inertia and B the transverse moment of inertia of the bullet about its centroid, and if N denotes the rate of spin, then the *stability factor* S of the projectile is given by $A^2N^2/4B\mu$. If S exceeds unity, the effect of the overturning moment is to cause the precession we have just described. If S is less than unity, the bullet will tumble and its flight will be erratic. However, μ is a function both of the velocity and of δ, so that an initially unstable bullet will lose velocity very rapidly, whereas the rate of loss of spin, for any bullet, is very small. Thus, the stability factor may increase in the first few feet of flight to sufficiently far above unity that sometimes an initially unstable bullet will not tumble, although such a bullet will not hit the target either, as the initial yaw will be large enough to produce cross-wind forces sufficient to divert the bullet far away from the initial trajectory.

In practice, S is made big enough by making N big enough. If, however, S is too large, the axis of spin will not tend to approach tangency to the trajectory, but will tend to remain fixed in space. Such a bullet, if fired at a high angle, may come down, still spinning, but tail first. If N is larger than necessary, the accuracy life of the weapon will be shortened.

For many years, the big problem of exterior ballistics was to find suitable expressions for the resistance **R** or **D**. Having found a suitable resistance function for a given projectile, taken as standard, the retardations of this projectile and of other presumably similar projectiles were found by dividing this function by the "ballistic coefficient" of each projectile with respect to the standard projectile, where the ballistic coefficient C is given by $C=m/id^2$. Here m is the mass (or, more usually, the weight) of the projectile, d is its diameter, and i is a constant called the "coefficient of form" (or, by the ballisticians of E. I. duPont de Nemours and Company, the "coefficient of ignorance"). It is an empirically determined constant which allows for the fact

that the projectiles being compared with the standard projectile are not *exactly* similar in all dynamic and geometric particulars. Of course, i is unity for the standard projectile.

In 1883, it was shown by Majewski that the resistance functions of some of the projectiles of his day could be expressed by the quantities given in the first part of Table I. Later work has

TABLE I. Expressions for the resistance R for various ranges of velocity v.

R	v (ft/sec)
M_1v^2/C	<790
M_2v^3/C	790–970
M_3v^6/C	970–1230
M_4v^3/C	1230–1370
M_5v^2/C	1370–2300
M_8v^2/C	1370–1800
$M_6v^{1.7}/C$	1800–2600
$M_7v^{1.55}/C$	2600–3600

changed the range of Majewski's last equation, and has added two more, given in the second part of the table.

For the past 15 or 20 years, attempts to express resistance by an analytic function have been abandoned. Today, resistance functions are determined by numerical integration from the firing data, and are given in the form of tables. Several of these functions have been determined by the Ballistic Research Laboratory, Aberdeen Proving Ground, and are known simply as G_1, G_2, $\cdots$. They are merely numerical values for the velocity gradient $-dv/ds$—a quantity that is more convenient, in many cases, than dv/dt. The retardations of military projectiles can be adequately represented by one or more of these functions, and suitable ballistic coefficients are determined empirically. The functions G_5 and G_6 are those generally used for small-arms bullets.

The Measurement of Velocity

The oldest apparatus for measuring velocities or projectiles is the ballistic pendulum, the theory of which is probably well known to every reader of this journal. Modern methods of measuring velocity involve the measurement of the time of flight of the projectile over a suitable base line. The average velocity over this base line is taken to be the instantaneous velocity at the midpoint (although the expression "instrumental velocity" is seen in some of the books).

The Le Boulengé chronograph.—The most widely used chronograph is that devised by M. Le Boulengé approximately 80 years ago. Essentially, it consists of two vertical rods suspended from electromagnets. In the circuit of the electromagnet supporting the longer rod, called the *chronometer*, is a length of No. 38 or No. 40 hard-drawn (not annealed) copper wire, stretched across the path of the bullet 3 ft from the muzzle of the gun. As the bullet travels down the range it breaks the wire, thus stopping the current to the electromagnet, and the latter drops the rod. At the far end of the range, the bullet strikes a thin plate hanging loosely from two hooks through holes near its top edge, and with its bottom edge bearing lightly against a sharp contact point. When the bullet strikes this plate, it passes through, leaving a ragged hole, but it also moves the plate away from the contact point. This breaks the circuit of the electromagnet holding the shorter rod, called the *registrar*. The registrar drops and strikes a trigger, which releases a knife that produces a mark on the still falling chronometer. From the position of this mark, one can determine the time of flight of the bullet from muzzle wire to terminal target. In order to keep the time interval between the impact of the bullet and the consequent breaking of the circuit as small as possible, it is necessary to have the bullet strike the lower half of the terminal plate.

Before using the instrument, it is adjusted as follows:

(1) With the chronometer held in place, the registrar is allowed to fall and thus produce a mark on the chronometer. This mark is called the "origin of coordinates."

(2) By means of a special switch, both rods are released simultaneously, and thus a mark is made on the chronometer in the usual way. The position of this mark is determined by the time required for the rods to fall and for the knife to be released and to travel the necessary (horizontal) distance to strike and dent the chronometer. This time interval is called the "time of disjunction" t_0, and the position of the mark on the chronometer is called the "height of disjunction" h_0, where $h_0 = \frac{1}{2}gt_0^2$. In most cases, the instrument is adjusted until t_0 is 0.015 sec or h_0 is 110.3 mm.

When the velocity of a bullet is measured, the position of the mark made on the chronometer is called the "height of record," although in many proof houses it is never recorded anywhere. It is determined by t_r, the sum of the time of

flight t_f and the time of disjunction t_0, so that $h_r=\frac{1}{2}gt_r^2$. Therefore,

$$t_f=t_r-t_0=\sqrt{(2/g)}(\sqrt{h_r}-\sqrt{h_0}).$$

The heights h_0 and h_r are measured with a special vernier scale to the nearest tenth millimeter, and the velocity is read directly from a table.

The electrostatic velocity system; measurement of retardation.—A bullet in flight carries a small electrostatic charge. This charge can be increased by means of a suitable brush discharge near the muzzle of the gun. The passage of the charged bullet through the plane of a loop of thin wire connected to a suitable amplifier can be used to produce an impulse, or signal, which will cause the chronometer to drop, and its passage through a second such loop will cause the registrar to drop. Or, if desired, the passage of the charged bullet through these two loops may be made to cause a pair of sparks to jump to the cylindrical surface of a synchronously driven rotating drum.

Recently, the Frankford Arsenal has used the apparatus originally devised for the aforedescribed hangfire test to measure the retardation of several types of small-arms bullet. Previous measurements of retardation involved firing through two or more screens, which, of course, affected the flight of the bullet, and it was seldom possible to fire through more than three screens at once. Therefore, one had to fire bullets over a few different ranges, and then try to reconcile a mass of data in order to obtain the resistance function.

At the arsenal, suitable loops and amplifiers were placed every hundred yards on the 600-yd range, and connected through a coaxial cable system to another amplifier in the firing room of the proof house; the output of this amplifier was in turn connected to the hangfire recorder. It was thus possible to time the transit of the bullet every hundred yards without interfering with the flight of the bullet. It was found that the retardation of most small-arms military bullets is given by $dv/dt=-kv^{\frac{1}{2}}$ for velocities between 2000 and 3000 ft/sec. The value of k for the 150-grain bullet known to hunters and target shooters as the caliber .30–06 is 0.0155 $ft^{-\frac{1}{2}}$ sec^{-1}.

In working up the data, it was found that the graph of the average velocity to each loop as a function of the distance to the loop was a straight line; that is, the bullets appear to obey the law

$$x/t=a-bx,$$

where x is the distance travelled by the bullet, t is the corresponding time of flight, a is a pseudo initial velocity (it would be the true muzzle velocity but for the fact that the retardation near the muzzle is greater because of the greater precession), and b is the slope of the straight line. Differentiating this expression twice, we obtain

$$d^2x/dt^2=-(2b/\sqrt{a})v^{3/2}=-kv^{3/2},$$

since the y- and z-components of the motion are negligibly small over 600 yd. Expressed in terms of the velocity gradient, the preceding equation becomes

$$-\frac{dv}{dx}=-\frac{1}{v}\frac{dv}{dt}=k\sqrt{v}.$$

On a range as short as 600 yd, $v^{\frac{1}{2}}$ is not distinguishable from Majewski's $v^{1.55}$, but it is easily distinguishable from his $v^{1.7}$. Re-examination of the Aberdeen G_5 function shows that for velocities between 1650 and 2950 ft/sec this function is equivalent to

$$\begin{aligned}-dv/ds&=G_5/C_5=0.118\sqrt{(v/1000)}/C_5\\&=0.00373\sqrt{v}/C_5;\end{aligned}$$

therefore $C_5=0.00373/k$.

To measure k, we need to fire through but three screens. Let x_1 and x_2 be the distances between the first and second and the first and third screens, respectively, and let t_1 and t_2 be the corresponding times of flight. Then

$$k=2(x_1t_2-x_2t_1)[x_1x_2t_1t_2(x_2-x_1)(t_2-t_1)]^{-\frac{1}{2}}.$$

The counter chronograph.—These chronographs contain a high frequency oscillator—10^5 or 10^6 cycle/sec. The passage of the projectile through the plane of a loop causes these oscillations to be fed to a counting circuit; the passage of the projectile through a second loop stops the counting. The number of whole oscillations counted by the apparatus is read directly from a lamp board operated by a scaling circuit. This number is the time of flight of the projectile between the two loops, either in microseconds or in hundredth milliseconds, depending on the frequency of the oscillator. Using the reading of the lamp board, the operator can obtain the velocity directly from a table.

The Accuracy Test

The accuracy of small-arms bullets is tested by firing them in a weapon having a heavy cylindrical barrel, with a wide bearing ring, integral with the barrel, at each end. The weapon lies in a steel trough or V block. The impulse produced by firing the shot causes the accuracy weapon to slide backwards in the V block. Supposedly, a barrel thus supported and allowed to recoil will not jump or whip when fired. However, care must be taken that the gun rests in exactly the same position each time it is fired, as the axis of the bore can never be *exactly* concentric with the bearing rings.

Ten shots are fired at a paper target, generally 500 or 600 yd away. After the shots are fired, the dispersion of the shot holes about their own centroid, or "center of impact," is determined. For many types of small-arms ammunition, the radial standard deviation of a ten-shot target at 600 yd is between 5 and 10 in.

As previously stated, the path of a precessing bullet is a spiral helix. This is true of a perfect bullet. However, manufactured bullets are not perfect: The cross sections will not be a perfect circle; the base will not be exactly normal to the axis of figure; the nose will not lie exactly on the axis of figure; the walls of the bullet jacket will not have exactly the same thickness all around; other components of the bullet will not be exactly homogeneous and uniform; in short, no bullet will be a perfect figure of revolution. These defects, including those usually called static unbalance in other fields, are grouped together under the term *dynamic unbalance*. Dynamic unbalance should not be confused with dynamic instability, described previously. Instruments for measuring dynamic unbalance of small-arms bullets have been devised by Professor Beams and his collaborators at the University of Virginia.

The *size* of a target—that is, the dispersion of the shot holes—in an accuracy test is a function of: (i) variations in muzzle velocity; (ii) variations in both magnitude and direction of initial yaw; (iii) variations in bullet weight; (iv) dynamic unbalance of the bullets; (v) variations in the position of the accuracy weapon during the test. Here we have ignored the wind conditions on the range.

The size of the target will also depend on the condition of the bore of the gun. It is found, at least for some guns, that guns which had been fired a few hundred times yielded smaller targets than did the same guns either when brand new or after having been fired a few thousand times. The internal condition of the gun affects chiefly the variations in initial velocity and yaw, so that the effect of the gun may be included in the first two variations mentioned.

The effects of the variations in velocity and in bullet weight are easily computed. The effects of the other variations are not so easy to calculate exactly.

Let the average initial velocity of a group of bullets be a, and let the standard deviation of this velocity be da. Let the position of the target below the line of departure be h, where $h=\frac{1}{2}gt^2$, and t is the time of flight of the bullet from muzzle to target. Let dh be the contribution to the vertical component of the size of the target, where $dh=gtdt$. Then, for small-arms bullets obeying the law

$$dv/dt=-kv^{\frac{1}{2}},$$

we have

$$t=2x/(2a-kx\sqrt{a}),$$
$$dt=-t^2(4\sqrt{a}-kx)da/4x\sqrt{a},$$
$$dh=gtdt=-gt^3(4\sqrt{a}-kx)da/4x\sqrt{a}.$$

For the service ball ammunition, $da=15$ ft/sec, $k=0.0155$ ft$^{-\frac{1}{2}}$ sec^{-1}, and $t=0.9$ sec when $x=1800$ ft and when $a=2775$ ft/sec. Then $dh=\frac{1}{6}$ ft $=2$ in. (approximately). But the vertical standard deviation of a target of service ammunition is generally 6 in. or so, whereas for specially manufactured match ammunition, the vertical standard deviation is frequently close to 4 in. at 600 yd. We may therefore conclude that the variations in muzzle velocity have negligible effect on the accuracy of service ammunition, but do have appreciable effect on the accuracy of match ammunition.

Similarly, for the effect of the variation in the weight of the bullets, we would have, for constant muzzle velocity,

$$dt=t^2\sqrt{a}dk/2=-t^2k\sqrt{a}dm/2m,$$

if we assume that k is inversely proportional to the mass of the bullet. Under present manu-

facturing conditions, dm/m is about 0.01, so that

$$dh = gtdt = -gt^3k\sqrt{a}dm/2m = 0.09 \text{ ft}$$
$$= 1 \text{ in. (approximately).}$$

It would thus appear that yaw and unbalance are the factors that contribute the most to the size of the accuracy target.

BIBLIOGRAPHY

Here are listed a few of the articles on ballistics (not necessarily of small-arms ammunition) that have appeared in journals familiar to physicists.

Bairstow, Fowler and Hartree, Proc. Roy. Soc. **A97**, 202 (1920).
Fowler, Gallop, Lock and Richmond, Phil. Trans. **221**, 295 (1920).
A. G. Webster, Proc. Nat. Acad. Sci. **548** (Nov. 1920).
J. Proudman, Proc. Roy. Soc. **100**, 289 (1921).
E. Bolle, Zeits. f. tech. Physik **3.6**, 205 (1922).
Fowler, Gallop, Lock and Richmond, Phil. Trans. **222**, 227, (1922).
J. C. Karcher, Sci. Pap. Bur. Stand. **445**, 257 (1922).
K. Rottgardt, Zeits. f. tech. Physik **4.2**, 63 (1923).
Gossot and Lioville, Comptes rendus **180**, 1014 (1925); **183**, 503 (1926).
C. A. Clemmow, Phil. Trans. **227**, 345 (1928).
von Kármán and Moore, Trans. ASME (June 1932).
Cranz and Schardin, Zeits. f. tech. Physik **13.3**, 124 (1932).
R. H. Kent, Mech. Eng. **54**, 641 (1932).
Crow and Grimshaw, Phil. Mag. **15**, 529 (1933).
Taylor and Macall, Proc. Roy. Soc. **139**, 278 (1933).
R. d'Adhemar, Ann. Soc. Sci. Bruxelles **57**, 73, 173 (1937).
R. H. Bacon, Phys. Rev. **64**, 44 (1943).
F. E. Grubbs, Ann. Math. Statistics **15**, 75 (1944).
Epstein and Churchman, Ann. Math. Statistics **15**, 90 (1944).

Of the following books, the first two are, today, chiefly of historical interest. Cranz's *Lehrbuch* is the most comprehensive treatise, consisting of four volumes; an annotated English translation of Volume II, "Interior ballistics," is in preparation, under the auspices of the National Defense Research Committee, by C. C. Bramble, H. Bluestone, J. D. Elder and D. Roller.

N. Majewski, *Traité de ballistique extérieure* (Paris, 1872).
James M. Ingalls, *Exterior ballistics* (Van Nostrand, 1886).
C. Cranz, *Lehrbuch der Ballistik* (Berlin, 1925).
F. R. Moulton, *New methods in exterior ballistics* (Chicago, 1926).
British textbook of small arms (Woolwich Arsenal, 1929).
E. E. Herrmann, *Exterior ballistics* (U. S. Naval Academy Institute, 1935).
T. J. Hayes, *Elements of ordnance* (Wiley, 1938).
W. Coxe and E. Beugless, *Exterior ballistic nomographs* (E. I. duPont de Nemours, 1938).
C. S. Robinson, *Thermodynamics of firearms* (McGraw-Hill, 1943).

THE industrial revolution has not only changed the nature of war as a technic, but it has also changed its political character. It is no longer a calculable instrument of politics, capable of achieving the aims for which it is waged, but escapes control and becomes a universal mechanism of disturbance not only for the belligerent nations but for all others as well. Once understood, this great historical fact must ultimately lead to the renunciation of war as an instrument of policy. Even the Hitlers of the future may hesitate to use it for their own advantage or that of a nation whose destinies they direct. There is no more compelling task at the present time than to make sure that this inevitable lesson of the nature of modern war is driven home in the minds of all thoughtful people in every land.—JAMES T. SHOTWELL.

Physics in the Navy*

FRED K. ELDER, JOHN A. TIEDEMAN, LAWRENCE E. KINSLER, JOHN D. RIGGIN, E. R. PINKSTON AND RALPH A. GOODWIN
United States Naval Academy, Annapolis, Maryland

INTRODUCTION†

TWO purposes are served in the presentation of this paper: (i) to show the widespread applications of physical principles in the life and the operation of the Navy; (ii) to show how a general physics course has been molded in order to train future naval officers in a thorough grasp of the fundamental concepts on which those applications are based. A selection, at random, from the routine duties and activities on board a ship will indicate the great diversity of uses to which physical principles are put in the Navy. For instance, every time the navigator takes a sight with his sextant he utilizes the principle that the rotation of a plane mirror upon which a ray of light falls causes the reflected ray to rotate also, the rotation of the reflected ray being twice that of the mirror. The listener on the sound watch is utilizing principles of diffraction and reflection of sound waves every time he takes a sound bearing. Every roll of the ship demonstrates the periodicity of simple harmonic motion. The wheel watch is using the principles of stability and precession of the gyroscope as he steers the ship on her course. The ship's engineering plant converts the energy of the fuel oil being fed to the boilers into kinetic energy. Every time a shell is hoisted from handling room to turret numerous electrical principles are put to use, to say nothing of such concepts as work, power and mechanical advantage. Gunnery involves the law of gravity, Newton's third law, the gas laws and the vector addition of velocities. Throughout the service many hundreds of such principles are put to work, and it is the intelligent use of these principles that helps to make the Navy effective.

The question arises, what is the minimum that the Naval Academy should offer to the future naval officer in the field of physics. The obvious answer to this question is—as much as time will permit. While the Physics Committee is, at present, a part of the Department of Electrical Engineering, it serves all of the professional and semiprofessional departments by teaching fundamentals and basic concepts. Much of the advanced work in physics, which is ordinarily taught in colleges by the physics department, is taught at the Naval Academy by other departments. In two topics, however, sound and light, the terminal study is made in the physics course, subsequent work therein being in the nature of refresher or drill.

Physics, then, is taught as a basic science course. The physical principles are applied in one form or another as engineering subjects in the Departments of Ordnance and Gunnery, Marine Engineering, and Seamanship and Navigation. For example, physics teaches the midshipman the various forms of motion and the trajectory; Ordnance and Gunnery shows how to use these concepts practically. Physics teaches the elements of heat, its measurement, transfer and conservation; Marine Engineering teaches the midshipman thermodynamics and gives him an insight into energy analysis of naval machinery. Physics starts the midshipman on his first work in electricity; Electrical Engineering takes him through a long and interesting practical course in electrical engineering. Physics teaches vectors; Seamanship and Navigation applies them to dead reckoning, surface piloting, aerial piloting and marine surveying. Physics teaches the fundamental concepts of work, energy and power, and the principles of machines; Seamanship and Navigation shows practical applications. Thus all of the fundamental concepts that the physics course teaches are used and elaborated in the other departments, but with a view to giving

* The opinions or assertions contained herein are the private ones of the writers and are not to be construed as official or reflecting the views of the Navy Department or of the naval service at large.

† By Fred K. Elder, Commander, USN (Retired), Head of Physics Committee, U. S. Naval Academy.

the midshipman a workable picture of the practical uses to which these principles are put.

From these comments it can be seen that the course in physics at the Naval Academy must be molded so as to cover all of the basic physical principles which give form to the work of all of the professional and semiprofessional departments. Moreover, the Physics Committee tries to keep abreast of the times, to follow up each new application to naval warfare, and to check forthwith the subject matter of the course, in order to insure that the midshipman is receiving a clear understanding of the basic concept upon which any new application is founded.

Naturally, there are many topics, some of which are secret, that cannot be discussed here. However, it is hoped that the present paper will indicate some of the multifarious applications of physics in the Navy.

ELEMENTARY VECTOR METHODS†

Although directed line segments whose length is proportional to either displacement or velocity have been used in the navigation textbooks of the Navy for years, there seems to have been a prejudice against calling them vectors. However, the latest edition of Dutton,[1] in the recently added chapters on aerial navigation, calls such line segments *vectors*, though the term has not yet crept into the older portions of the book.

Students who have had the type of elementary physics course designed for engineers and physicists will have little difficulty in solving naval problems involving vectors. Since many such problems involve only the concepts of displacement, time and velocity, it is desirable to introduce these concepts early in the physics course as a means of providing practice in dealing with vectors, at a time when the student is not prepared to deal with force and acceleration.

Types of problems that are most useful early in the physics course are those known in the Navy as "maneuvering board problems."[2] One of the simplest problems involves two ships—a guide ship which is proceeding on a given course with a stated speed, and another ship which is sailing with the same course and speed at a given distance from the guide and on a given bearing. The second ship is ordered to proceed to a new point with reference to the guide. The maximum speed of the second ship is given, and it is necessary to find the course to be steered and the time necessary to complete the maneuver. In addition, it is usual to find the bearing and the distance to the guide at several times during the maneuver, as a check on the accuracy with which the maneuver is carried out.

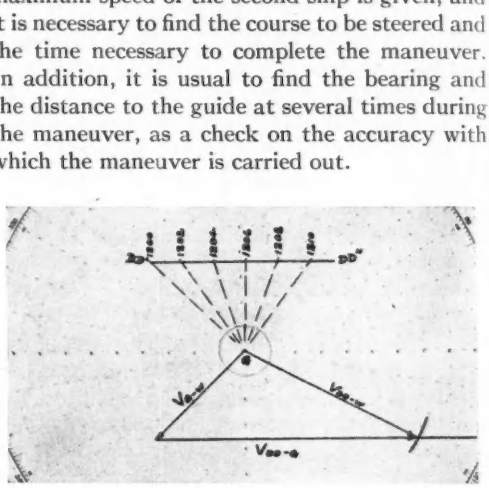


FIG. 1. A simple maneuvering board problem.

This is a very practical problem, since capital ships are always accompanied by a screen of lighter vessels. The detachment of one of the latter to investigate reports of the presence of enemy submarines makes it necessary to rearrange the screen. Since all ships are equipped with peloruses, which enable the captain to obtain rapid and accurate bearings of distant objects, and nearly all are equipped with range finders, the information for carrying out the maneuver is available.

Problem 1. A capital ship is proceeding on course 225° T, speed 20 knots, with a destroyer on the starboard beam 5 mi distant. The destroyer is ordered at 1200 to take a position dead astern, distant 5 mi. Find: (a) course of the destroyer at 30 knots; (b) time required to complete the maneuver; (c) predicted bearing and distance at 2-min intervals.

To conform to the nautical convention, the capital ship is labelled G (guide), and the destroyer is labelled DD' in its first position, and DD'' in its second (Fig. 1); DD' and DD'' are positions relative to the guide. Clearly the

† By John A. Tiedeman, Lieutenant Commander, USNR.

[1] Dutton, *Navigation and nautical astronomy* (U. S. Naval Institute, Annapolis, 1944).

[2] Many problems of this type are worked in reference 1; also in U. S. Navy Hydrographic Office, *No. 217 Maneuvering board manual* (Government Printing Office, 1941).

displacement and velocity of the destroyer relative to the guide are in the direction $DD'-DD''$. The velocity of G relative to the water is 20 knots in the direction 225° T. We can use the vector equation

$$\mathbf{V}_{DD-W}=\mathbf{V}_{DD-G}+\mathbf{V}_{G-W}, \qquad (1)$$

where $\mathbf{V}_{DD-W}$ means "velocity of destroyer relative to the water" with similar meanings for the other symbols. The first term in Eq. (1) is unknown in direction, but is 30 knots in magnitude, while the second is known in direction only. The solution is simple graphically. Plot G, DD' and DD''. Draw the vector $\mathbf{V}_{G-W}$. From its tip, draw a line of undetermined length parallel to $DD'-DD''$. From G strike an arc of length 30 units. The vector connecting G with the intersection of the arc and the line parallel to $DD'-DD''$ is $\mathbf{V}_{DD-W}$, the velocity of the destroyer relative to the water. Since the relative displacement can be measured, we can determine the time required to complete the maneuver by dividing the displacement by the relative velocity.

It is a little difficult for the student to grasp the fact that the problem actually involves two diagrams, one dealing with displacements and having a vector scale in nautical miles, the other dealing with velocities and having a scale in knots. Returning to the displacement portion of the diagram, we can locate the destroyer relative to the guide at any time by multiplying the elapsed time and the relative velocity. Doing this for intervals of 2 min, we obtain the guide positions *1*, *2*, *3*, $\cdots$. The length and orientation of the lines connecting these points with the guide give the bearings of the guide and its distance at 2-min intervals.

In carrying out such a maneuver, the navigator works the problem in the afore-described manner. The captain or officer of the deck gives the order to the helmsman to steer the course given by the navigator at the proper time. He instructs his quartermaster to obtain bearings and distances of the guide at 2-min intervals. The quartermaster's reports are then checked with the navigator's prediction.

The advance and transfer of all ships of the Navy are known. The *advance* is the distance progressed on the original course when a turn is made, and the *transfer* is the distance traveled normal to the original course. When maneuvers like the one described are made with ships that are close together, advance and transfer must be taken into account by the navigator.

Of course, many maneuvering problems are far more complicated than the one just discussed. As examples, there is the problem of what course to take in intercepting the enemy; or of how to pass an enemy vessel beyond the limit of visibility, or beyond range; or of the course an auxiliary ship should take to remain in the zone of protection afforded by capital ships for the maximum time; or of scouting on a given bearing from the guide and returning within a fixed time. These problems are more complicated than the simple one discussed, but are solved by vector methods familiar to all physicists.

As a matter of convenience, such problems are worked on specially ruled paper called a *maneuvering board*. This contains a series of concentric circles, with radial lines spaced 10° apart. At the bottom (not shown in Fig. 1) there is a nomograph for the solution of problems involving time, velocity and distance. By marking a known quantity on each of two scales—velocity and distance, for example—and connecting the two with a straightedge, the third quantity—say, time—may be read on the third scale.

Another type of vector problem involves the motion of a ship in a moving medium, for example, an airplane flying when there is a wind. Similar problems arise in the navigation of surface ships, though ocean currents rarely have a high velocity. Problems involving the wind relative to a moving ship are important in the take-off and landing of airplanes on carriers.

Problem 2. An airplane wishes to make good a course of 135°, with a 60-knot wind blowing from 180°. The plane is capable of an air speed of 240 knots.

For this problem Eq. (1) becomes:

Velocity of plane relative to the air+velocity of air relative to earth=velocity of plane relative to earth.

For the first term, only the magnitude is known; the second is completely known; for the third term, only the direction is given. From the origin O draw the wind vector **OW** and also the course to be made good, a line of undetermined length. From the tip W of the wind vector, strike an arc of length 240 units so as to cut at P the line representing the course to be made good. The distance OP represents the speed made good, while the direction WP is the course to be steered.

Vector problems similar to Problem 2 arise so frequently in the navigation of aircraft that all naval aircraft are

supplied with an aircraft navigational plotting board, which is similar to the maneuvering board in purpose, but has a celluloid working surface from which marks can be erased. Hence it may be reused many times.

Naval ordnance problems are more complex than maneuvering and navigation problems, since the projectile is in the air (which may be moving) for many seconds. Before firing, an estimate of the course and speed of the enemy must be made. The range and bearing of the enemy must be found, since they determine, respectively, the elevation of the gun and the direction in which it is trained. Since both ships ordinarily are moving, there are continuous changes in both elevation and train. Even the construction of range tables is based on an application of vectors, as is that of the tables giving the numerous corrections which must be made. The actual problem of how to elevate and train the gun could not be solved rapidly enough in these days of high speeds without the aid of mechanical devices, which, when set on the basis of observed bearing and speed of target, give an accurate and continuous solution of the problem.

The application of mechanical devices to the solution of ordnance problems has made possible enormous improvement in accuracy of gun fire. During the battle of Manila, in the Spanish American War, it was estimated that 2 percent of the shots fired by us were hits. Fortunately, the Spaniards were not even that good. Great improvement followed after the adoption of range finders, until now it is not uncommon that the first salvo fired scores a hit.

Although it is not permissible to describe the mechanical devices used, the study of problems involving relative motion, particularly the components of relative motion resolved parallel and perpendicular to the line of sight between two moving vessels, will with certainty lay an excellent foundation for the understanding of all ordnance gear.

The application of vectors to the stresses in structures is as important in the training of the naval officer as in the training of the average engineer. Derricks must be frequently rigged, and cargo and loose gear must be blocked or lashed to withstand considerable rolling and pitching.

There are many special problems in the Navy to which vectors are applicable, in a general way, but in which the quantities involved are not well enough known for an exact solution. Among these are the problems of seamanship, which include towing, mooring, stranding and general ship handling, as well as problems of cargo handling and the rapid repair of gear damaged at sea or by enemy action. Until recently these problems were treated entirely on a "rule of thumb" basis, and little attention was directed to the similarity between them and the ones studied by midshipmen in physics, mechanics and marine engineering. Vector problems involving forces, worked in any physics course, give the student a background which enables him to supply reasons for the old "rule of thumb" methods, so that it is easier for him to master or to recall them.

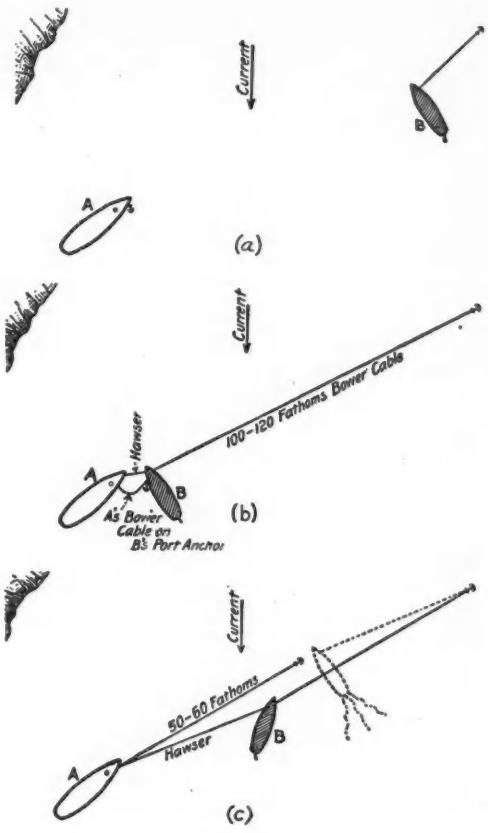


FIG. 2. One method of assisting a stranded vessel.

For example, one of the problems of seamanship is assistance to a stranded vessel. One method requires the assisting vessel B to drop an anchor at a distance, as in Fig. 2(a), ease off toward the stranded vessel A, and take off her anchor and two hawsers. The anchor is carried out as shown in Fig. 2(c). To quote Knight:[3]

> [Ship] B heaves in on her steam windlass and goes ahead with her screw. Thus we have the windlasses of both ships pulling on A, with the power of B's screw added, and still further the "sucking effect" of current acting on B's port bow, and providing A's bow yields enough to cant her, acting on A as well.

Everything said about the problem is sound physics, and the student who can visualize a force applied to A's anchor, the strain in the hawser, the current carrying B downstream, and the strains taken in the hawsers by winches on both A and B can readily apply the method. If the student has had drill in the vector solution of problems involving a weight hanging on a taut horizontal line, he should see implications not suggested by Knight. Experience with the vector solution of force problems makes it possible to see applications in new and unfamiliar fields.

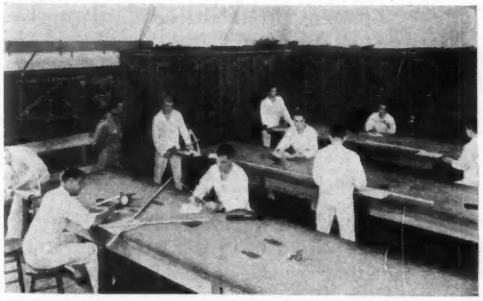
MECHANICS†

The basic principles of mechanics are of course indispensable to an understanding of the operation of naval ships and airplanes. There is probably no principle, law or theory that lacks at least some application in either the development or operation of such complicated engineering plants as are contained in the modern warship or military airplane. Articles similar to the preceding one on vectors and the following ones on gyroscopes could be written on such subdivisions of mechanics as accelerated motions, trajectories, force, equilibrium, machines, energy and power, hydrostatics and hydrodynamics. Some of these topics are quite adequately treated in first-year college physics textbooks while others are not.

The following topics are ones which should be comprehensively treated in a mechanics course as being of particular importance in studying for a naval career: (1) equations of motion with constant velocity and with constant acceleration, the emphasis being on trajectories of projectiles; (2) equilibrium and stability of rigid bodies; (3) simple machines and basic mechanisms; (4) Newton's laws of motion; (5) hydrostatic phenomena; (6) momentum and energy; (7) gyroscopic motion.

A complete understanding of the equations and of the various elements of the trajectory is most important since the ultimate function of naval ships and airplanes is destruction of the enemy through accurate projection of explosive missiles. Intricate devices such as rangekeepers and bombsights have been developed to facilitate this operation. However, in order to understand the theory behind their construction and to appreciate the relative importance of the various factors considered in their use, a knowledge of trajectories is essential. Furthermore, in the event of failure of our complex mechanical aids it becomes necessary to fall back on simple "rules of thumb" or graphical methods based on fundamental principles.

The beginning course in physics should include a discussion of the primary equations for a trajectory in vacuum. Although air resistance modifies their predictions, these equations constitute the starting point of the treatments found in training pamphlets, technical papers and books dealing with bombing, fire control and exterior ballistics. It is customary to superimpose the effects of air resistance as a correction to the predictions in vacuum.



Official U. S. Navy photograph

Laboratory experiment on ballistics.

† By Lawrence E. Kinsler, Lieutenant Commander, USNR.

The parametric equations of motion in vacuum are

$$v_x = V \cos \phi, \tag{1}$$

$$v_y = V \sin \phi - gt, \tag{2}$$

$$x = V \cos \phi t, \tag{3}$$

$$y = V \sin \phi t - \tfrac{1}{2} g t^2, \tag{4}$$

where V is the initial velocity and ϕ is the angle of departure as measured in a vertical plane between the horizontal and the line of departure.

The ability of a student to use these general equations will facilitate his understanding of the numerous special cases that will be encountered. Important cases resulting from the introduction of specific initial or final conditions are: (1) horizontal bombing, (2) dive bombing, (3) vertical bombing, (4) surface gunnery, and (5) elevated gunnery.

In *horizontal bombing*, V is the velocity of the airplane, $\phi = 0$ and $y = -h$, where h is the altitude above the target. This is the method used in high altitude bombing by heavy and medium bombers, the dropping of depth charges by aircraft participating in antisubmarine warfare, the release of torpedoes by torpedo bombers and "skip bombing" by low flying bombers.

Dive bombing, which differs from horizontal bombing in that ϕ is negative, is effective in operations against small targets and in countering evasive tactics of moving targets.

In *vertical bombing*, $V = 0$ and $y = -h$, and the resulting equations are those for a body dropped from rest. Typical examples are the bombing of submarines from stationary blimps, the projection of zero-velocity bombs by airplanes, and the bombing of one airplane by another flying overhead at the same speed.

In *surface gunnery*, V is the muzzle velocity of the projectile, ϕ is positive and $y = 0$. This type of trajectory is characteristic of ship-to-ship gunfire and of the bombardment of land targets at sea level.

Elevated gunnery differs from surface gunnery in that $y = h$, where h is the altitude of the target above the earth's surface. This type of gunfire is encountered in antiaircraft batteries, in the bombardment of land targets above sea level, and in star shell illumination of targets at night.

In solving trajectory problems certain quantities are of special importance. The *horizontal range* X, or distance $v_x T$ measured parallel to the earth's surface from the point of fire or bomb release to the point of fall, determines the angle of departure ϕ to which a gun must be elevated in order to hit a target. In bombing it determines the distance ahead of a target that a bomber must release its bomb. The *dropping angle* $\theta [= \tan^{-1} X/h]$ is the angle included between the vertical h and the line of sight s to the target at the instant of releasing a bomb (Fig. 3). It is the chief element in the mechanical solution supplied by a bombsight in horizontal bombing.

Knowledge of the total *time of flight* T of a projectile has many uses. When a ship is under bomber attack, T determines the time available for maneuvering the ship out of its original course. It determines the change in deflection, elevation and range for which correction must be made in allowing for the target's motion during the flight of the projectile. The fuze setting for fuzed projectiles depends upon T. The *maximum ordinate* y_0 delimits the types of antiaircraft weapons to be used in firing on enemy airplanes. Obviously, guns firing projectiles with a maximum ordinate less than the altitude of an attacking airplane cannot physically damage the airplane but serve merely to discourage flying at a lower altitude.

Knowledge of both the magnitude and direction of a projectile's *striking velocity* $v_\omega [= (v_x^2 + v_y^2)^{\frac{1}{2}}]$ is of great importance. A projectile's striking speed determines its ability to penetrate armor. If a torpedo released from an airplane strikes the water at an excessive speed, it will be damaged and not run true. A useful "rule of thumb" for pilots is that in bombing from below 500 ft, the initial velocity of the airplane largely determines the final speed of the projectile. Variations of altitude within this range will only slightly affect the final speed. Reverse conditions apply to bombing from altitudes in excess of 5000 ft. The *angle of fall* ω of a torpedo influences the initial depth reached and its tendency to ricochet or broach. Most ordinary projectiles will ricochet if their angle of fall upon striking the water is less than about 12°. Increasing ω decreases the *hitting space* behind a target. The latter is the distance behind a target by which a shot that normally strikes the top of the target would instead strike the base of the target if the target were displaced by this amount. Thus, a projectile with a flat trajectory resulting in a small angle of fall will hit a target even if the estimated range has a considerable error, whereas a projectile with a large angle of fall will not

strike the target if the estimate of range is slightly in error. It is for this reason that the maximum angle of elevation of main battery guns is limited to about 30° and that any future increases in range that may result from increased muzzle velocities will be more valuable than those resulting from increased angles of departure.

Four examples will be considered in detail as a means of illustrating typical trajectory problems for the beginner in physics. These problems are examples of those that will be later encountered in textbooks and training pamphlets used in instructing naval personnel. The principal aim of such instruction is to promote an understanding of and ability to use such mechanical devices as rangekeepers, bombsights and automatic gunsights, with their associated tables

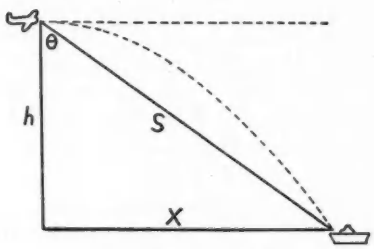


FIG. 3. Dropping angle θ.

and charts. It is essential to understand the fundamental principles of such devices in order to appreciate their limitations and to improvise should they fail.

Example 1. A heavy bomber is making a horizontal bombing attack at an altitude of 20,000 ft and a speed of 200 knots.

Here $h=-20{,}000$ ft, $\phi=0$ and $v=200$ knots $=338$ ft/sec. Therefore, in view of Eqs. (1) to (4), $T=35.2$ sec, $v_x=338$ ft/sec, $v_y=-1130$ ft/sec, $v_\omega=1180$ ft/sec, $X=11{,}900$ ft, $\theta=30.8°$ and $S=7730$ yd. These results show that 35+ sec is available for the target ship to turn out of its course after the bombs have been released. The final speed of 1180 ft/sec determines the number of inches of armored deck that an armor-piercing bomb will penetrate. The dropping angle of 30.8° is the main element introduced into the bombsight solution. The complement of this angle, 59.2°, is the limiting angle of elevation below which the bomber must be shot down if it is to be prevented from releasing bombs aimed to hit the ship. The slant range S determines the types of antiaircraft guns that may be effectively used. The value 7700 yd indicates that this bomber must be shot down by large caliber antiaircraft guns or friendly fighter planes if it is to be prevented from bombing the ship.

Example 2. An airplane flying at 100 ft elevation and 150 knots releases a depth bomb aimed to sink a submerging submarine.

Here $h=-100$ ft, $\phi=0$ and $V=254$ ft/sec. Therefore, $T=2.5$ sec, $X=212$ yd, $v_x=254$ ft/sec, $v_y=-80.5$ ft/sec and $\omega=17.6°$. Since the horizontal range is 212 yd the pilot must release his depth bomb more than 200 yd in advance of the point at which the bomb is intended to strike the water. The angle of fall of 17.6° indicates that the bomb probably will ricochet and thus upset the calculated position and time of its entry into the water. This trajectory problem is further complicated by the necessity of considering the forward motion of the submarine during the 2.5 sec of flight of the bomb plus the time required for the latter to sink to the set depth of explosion. Finally, the depth bomb must explode within a few feet of the submarine in order to be lethal. Consequently, when a pilot "sights sub and sinks same" he has applied physical equations in a manner that warrants the top naval grade of 4.0.

Example 3. A battleship fires a projectile of muzzle velocity 2600 ft/sec at another ship 45,000 yd distant.

Here $X=135{,}000$ ft, $y=0$ and $V=2600$ ft/sec. Therefore, $\phi=20°$ and $T=55.2$ sec. If the two ships are closing toward each other at a range rate of 20 yd/sec, the advance range to be used in order to hit the ship would be reduced from 45,000 yd to $45{,}000-(55\times20)=43{,}900$ yd. Correspondingly, the value of the angle ϕ will have to be made less than 20°. Similarly, from the deflection rate, the time of flight and the range it is possible to calculate the deflection angle by which the angle of train of the gun must lead the line of sight. Assuming a deflection rate of 20 yd/sec at right angles to the line of sight, the lateral displacement of the target ship is $55\times20=1100$ yd. This displacement subtends at the firing ship an angle of approximately $1100/45{,}000=0.0245$ rad, or 24.5 mils. Therefore, the correct deflection angle of lead is 24.5 mils. Many additional complicating factors such as the effects of air resistance, wind and drift caused by gyroscopic precession may be taken into account by means of range tables, once the values of X and T are known.

Example 4. The airplane of Example 2 releases a torpedo instead of a depth bomb.

If the torpedo is to enter the water without damage caused by excessive deceleration—in excess of 100 g—its longitudinal axis at the instant of impact must be nearly tangent to the trajectory. Basic principles of mechanics may be utilized in any one of three different ways in order to bring this about. If the effects of air resistance were negligible, the torpedo could be released nose down at an angle of 17.6° while the plane was flying horizontally; because of its inertia the torpedo would enter the water nose down at this angle and thus tangent to the trajectory. Again neglecting air resistance, the torpedo might be released with its longitudinal axis horizontal, but with a nose-downward angular velocity of 7°/sec about a lateral axis through its center of gravity; since no torque is acting about this axis it would continue to rotate at this rate

during the 2.5 sec of fall and therefore would assume an angle of 17.5° upon entering the water. Finally, if the tail of the torpedo is equipped with stabilizers, aerodynamic forces can be utilized to bring the longitudinal axis of the torpedo tangent to the trajectory. If the center of pressure of the aerodynamic forces is aft of the center of gravity of the torpedo, it will be aerodynamically stable. Any torques that may exist will accelerate the torpedo about its center of gravity so as to bring the longitudinal axis parallel to the relative wind which in still air will be along the trajectory.

This treatment of trajectory problems may seem rather extensive to be advocated for an introductory physics course. Nevertheless, it should show that a complete understanding of the basic equations of motion is essential to a career in the Navy. Furthermore, if the beginner is trained in solving trajectory problems he should experience little difficulty in solving such acceleration problems as the catapulting of airplanes, the acceleration and deceleration of airplanes on carrier flight decks, the acceleration and deceleration of ships, and the acceleration of projectiles in gun barrels.

The principles involved in equilibrium and stability conditions are of particular importance. This includes not only the application of the conditions of equilibrium to crane booms, gun mounts and structural members but as well to the entire ship or airplane. The concepts of center of buoyancy, center of pressure and center of resistance are utilized concurrently with the more common ones of centers of mass and of gravity. Dynamic stability must be considered in addition to static or initial stability.

A consideration of the reaction forces on the gun mount of an antiaircraft gun provides a simple example of equilibrium. If the mount is bolted to the deck, the forces acting during recoil may be represented as in Fig. 4. Here R_r is the force of recoil exerted by the trunnions on the mount, R_1 is the reaction to the force tending to push the breech side of the mount into the deck, R_2 is the reaction to the force tending to tear the bolts at the muzzle side of the mount out of the deck, R_s is the reaction to the shearing force tending to shear off the bolts in a horizontal plane, W is the weight of the entire mount, including the gun. The forces R_1, R_2 and R_s may be determined by application of the usual conditions of equilibrium,

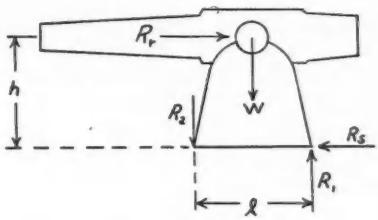


FIG. 4. Equilibrium of a gun mount.

$$R_r - R_s = 0, \quad R_1 - R_2 - W = 0,$$
$$R_r h - R_2 l - W \cdot \tfrac{1}{2} l = 0.$$

The results are used in designing the strength of the mount and of the deck structure on which the mount is placed. When the gun is elevated the force R_r can be resolved into a vertically downward component and a horizontal component acting similarly to R_r above, but of smaller magnitude. As a result, R_2 and R_s are decreased and R_1 is increased.

As another example consider the equilibrium and stability of a ship with respect to roll about a longitudinal axis. A ship floating at rest in an upright position is in equilibrium. The forces acting are the weight of the ship and the upward hydrostatic force on its wetted surface, and these two forces must be equal in magnitude. Further, since the total torque about any axis must be zero, the center of buoyancy and the center of gravity of the ship must lie in the same vertical line. If a ship is to possess static stability as it rolls, the center of buoyancy of the displaced water, which lies below the center of gravity, must be displaced sufficiently to result in a restoring torque about a longitudinal axis through the center of gravity. This is true when the metacenter M is above the center of gravity (Fig. 5). Under these conditions the restoring torque is $W \cdot GM \sin\theta$, where θ is the angle of heel, and the ship will be righted. Thus the *metacentric height* GM is a measure of a ship's static stability.

Suppose that the ship's hull is damaged so that one of the amidships starboard compartments is flooded. The weight of the water in this compartment affects the trim of the ship in two manners. The ship sinks deeper in the water until the buoyant force equals the new weight. As the reserve buoyancy of most high sided war-

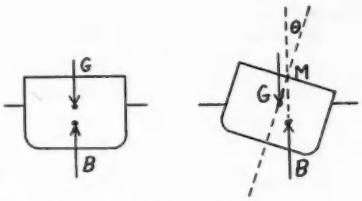


FIG. 5. Equilibrium and stability forces on a ship.

ships is 50 to 75 percent of normal displacement, the chances are that such a ship will not sink owing to loss of buoyancy. However, the ship will not be in equilibrium until it heels over and assumes a list to starboard such that the restoring torque caused by the new location of the center of buoyancy counterbalances the torque due to the weight w of water in the flooded compartment. This angle of list is given by the equation $W \cdot GM \sin \theta = wl \cos \theta$, where l is the distance from the center of gravity of the undamaged ship to that of the water in the flooded compartment measured in a horizontal direction when the ship is upright. This list to starboard can be countered either by pumping out the flooded compartment or, that being impractical, by counterflooding a compartment, tank or double bottom on the port side. This latter means will remove the list but will still further reduce the available reserve buoyancy.

The metacentric height GM is not constant for all angles of heel. As θ increases beyond 10° the value of GM begins to decrease. As a result the restoring torque does not increase directly as $\sin \theta$ but reaches a maximum between 30° and 50° and returns to zero between 60° and 90°. This latter condition is reached when the metacenter coincides with the center of gravity, resulting in a zero value for GM. A battleship will capsize if its angle of heel exceeds about 60°.

A complete discussion of the inclination of a ship requires a consideration of the energies involved. The *dynamical stability* of a ship is dependent on the amount of work done in inclining the ship to any angle of heel. It is measured by the product of the weight W of the ship and the change in the vertical distance between the center of gravity and the center of buoyancy. If a force capable of producing an inclination acts upon a ship, the inclination reached by the ship, before starting to return to the upright position, is not determined by the static moment of the inclining force, but by an equality of energies. The ship will heel over until the work done is equal to the energy initially associated with the inclining force. The angle of heel determined in this manner is, in general, larger than that for static equilibrium. If an inclining agent with energy greater than the dynamical stability at the limiting angle of static stability acts upon the ship, she will capsize even though the static moment of its force is less than the maximum righting torque of the ship. This explains why a sailing boat will frequently capsize when suddenly struck by a squall and yet will sail in perfect safety in a steady wind of higher velocity.

The operation of submerging a submarine (Fig. 6) offers many examples of the application of equilibrium conditions. While a submarine is running on the surface, its stability characteristics are similar to those of a surface ship. In submerging, a submarine must take into her main ballast tanks a weight of water exactly equal to the weight of water to be displaced by the conning tower and other parts of the ship normally above water. This requires that her total weight must not change radically during a cruise, for otherwise the weight of water taken into the main ballast tanks will not be sufficient to enable her to submerge. The loss of weight resulting from the consumption of such expendable supplies as fuel oil must be compensated for by taking water into the fuel tanks and variable ballast tanks.

When a submarine is submerged, the problems of equilibrium and stability are radically changed to the reverse of surface conditions. The center of buoyancy shifts upward and coincides with the metacenter. This means that to be stable a submerged submarine must have its center of gravity below its center of buoyancy. Such a submarine may be ballasted so as to have a range of stability of heel of 180°. Upon submerging, a submarine loses the large righting moments, and therefore stability, that all surface ships have toward pitching about an athwartships axis through their centers of gravity. A submarine maintains the trim of its longitudinal axis by

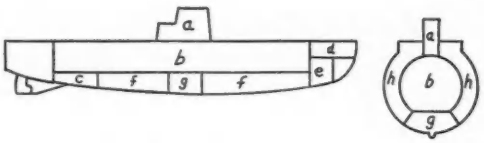


FIG. 6. Buoyancy and trimming compartments of a submarine: *a*, conning tower and bridge; *b*, living spaces; *c*, stern trimming tanks; *d*, bow buoyancy tank; *e*, bow trimming tank; *f*, fuel oil tank; *g*, safety tank; *h*, main ballast tank.

controlling the amount of water in bow and stern trimming tanks. In addition, it is necessary to provide adjustable horizontal diving planes to assist in adjusting the trim and depth of the submarine when it is in motion. A submarine tends to lose trim upon firing a torpedo. This is countered by flooding the torpedo tube until the flooding of bow trimming tanks, or shifting of spare torpedoes, provides a permanent remedy. An analysis of the angles of trim taken by a submerged submarine under various conditions of weight distribution can be shown to be analogous to those taken by a weighted meter stick supported above its center of gravity. Thus the operation of submarines requires a complete understanding not only of Archimedes' principle but of conditions of equilibrium and stability as well. Should the skipper of a submarine allow the center of gravity of his ship to approach too close to its center of buoyancy, it is entirely possible that the motion of one man from one part of the ship to another could seriously affect the vessel's trim.

Numerous important applications of the conditions of equilibrium and stability must be considered in discussing the flight of airplanes.

By considering the roll of a ship to be similar to the oscillations of a physical pendulum, it is possible to show that the *period of roll* T is

$$T=2\pi\sqrt{(k^2/GM\cdot g)},$$

where k is the radius of gyration of the ship about a longitudinal axis through G. Average values of this period range from 8 sec for submerged submarines to 20 sec for battleships. A knowledge of the period of roll of his ship and its significance is of importance to every commanding officer. If a ship is damaged and some parts flooded, the value of GM will be reduced and the period of roll increased. If it is observed that the period is doubled, this indicates that the metacentric height GM has been reduced to one-quarter of its normal value. Such a reduction in GM indicates a corresponding reduction in the restoring torques tending to right the ship. The range of stability would also be greatly reduced. Thus observation of the period of roll of a damaged ship will, in a matter of seconds, give her commanding officer a good indication as to the advisability of ordering "abandon ship" or not.

If the apparent period of the waves is equal to the natural period of a ship, the energy steadily imparted to the ship will cause it to roll through larger and larger amplitudes with resulting discomfort to the crew and danger of capsizing. The *apparent period* T_A of the waves is given by the Doppler equation,

$$T_A=T_W\frac{V_W}{V_W-V_S\cos\theta},$$

where T_W is the true period of the waves, V_W is the true velocity of the waves, V_S is the velocity of the ship, and θ is the angle between the course of the ship and the true direction of the waves. Therefore, synchronism between ship and waves can either be established or destroyed by a change of course or speed. Ships damaged in actions in the South Pacific have been assisted by changing course or speed when it was found that their periods of roll were in dangerously close synchronism with the apparent period of the waves.

Many uses are made of hydrostatic and other pressure phenomena. The linear relationship between pressure and depth in a liquid is used in calibrating pressure gages as depth gages for installation in submarines. Depth charges explode because the pressure at the set depth becomes sufficient to trip the firing mechanism. The depth setting on the pistol scale of a depth charge is adjusted by increasing the tension on a spring attached to a diaphragm. As the depth charge sinks, the hydrostatic pressure on the diaphragm steadily increases until the resulting force is equal to the set tensile force of the spring. When these two forces balance, the firing mechanism explodes the depth charge.

The force on the diaphragm is proportional to the hydrostatic pressure, this pressure is proportional to the depth, and the tension in the spring is proportional to the stretch. Because of these three linear proportionalities, the depth of explosion is directly proportional to the stretch of the spring, and this results in a linear scale of depth settings on the pistol.

Hydrostatic pressure is the source of one of the two physical forces actuating the depth control mechanism of torpedoes. If the force on a hydrostatic diaphragm is equal to the opposing force in a spring set at a tension corresponding to the desired running depth, the horizontal rudder will be in a neutral position. The movement of the diaphragm, when the torpedo is not at the set depth, causes a rudder angle to be put on the horizontal rudder so as to return the torpedo to this depth. As the torpedo could be running at the correct depth but diving deeper because of a nose-down trim, it is also necessary to use the force of gravity to actuate a trim-controlling pendulum. This pendulum puts a rudder angle on the horizontal rudder unless the torpedo axis is horizontal.

A knowledge of pressure magnitudes is essential to the operation of submarines. There is the obvious limitation as to safe depths because of increasing pressure and resulting total force that the hull must withstand. Modern submarines are of double-hull construction. The inner, or "pressure," hull is roughly circular in cross section. This hull must be constructed to withstand the extreme water pressure when submerged to extreme depths. The outer hull is a ship-shaped hull and gives the desired lines for surface operations, propulsion and seaworthiness. The space between the two hulls is used for ballast tanks, auxiliary tanks and fuel tanks. When submerged, these tanks are always open to sea pressure. As the outer hull is thus subjected to equal internal and external pressures, it can be made very light.

Upon submerging a submarine has some 500 ft^3 of air under a pressure of 3000 $lb/in.^2$ stored in various air flasks. An understanding of Boyle's law is essential to an economical use of this air.

Example 1. A submarine with main ballast tanks of volume 8000 ft^3 is submerged to a depth corresponding to a pressure of 150 $lb/in.^2$. What volume of compressed air will be needed to blow the ballast tanks at this pressure?

From Boyle's law, $3000V_1 = 150 \times 8000$, or $V_1 = 400$ ft^3. By this method almost the entire supply of high pressure air would be used. On the other hand, if allowance is made for the fact that as the submarine approaches the surface the pressure in the ballast tanks need not exceed atmospheric pressure, a much smaller volume is used, for then $3000V_1 = 15 \times 8000$, or $V_1 = 40$ ft^3. Compressed air is used on submarines not only to blow tanks but also as a means of storing large amounts of energy in a small space. It operates compressed air machinery for the operation of flood and vent valves. It is used to charge the air flask in torpedoes and thereby supply them with power for their run.

Example 2. If 20 ft^3 of air under a pressure of 3000 $lb/in.^2$ is driving a 100-hp turbine and exhausting at a pressure of 300 $lb/in.^2$, how long will the torpedo run?

Assuming isothermal expansion and 100 percent efficiency, we find that the energy available is

$$E = P_0V_0 \ln (P_0/P_1) = 3000 \times 144 \times 20 \ln (3000/300)$$
$$= 1.98 \times 10^7 \text{ ft lb.}$$

Therefore, since $E = Pt$, $t = 6$ min. If this torpedo has a speed of 30 knots, its range will be 6000 yd. The expansion actually would be more nearly adiabatic than isothermal and would result in a decreased running time and range. Consequently, the modern torpedo uses fuel to heat the air and supply additional energy.

An ability to estimate both speed ratios and force ratios of simple machines is important. Much heavy equipment must be moved on board ship. The prime movers usually are electric motors, steam engines, hydraulic devices such as the Waterbury variable-speed gear, or manually driven devices. The gear that must be handled includes 50-ft motor launches that must be hoisted aboard; 10-ton bower anchors that must be weighed; 1-ton projectiles that must be hoisted into the turrets from shell storage space; turrets, directors, rangefinders, rudders and boat cranes that must be rotated; airplane elevators on carriers that must be raised; and guns to be elevated. In the majority of these cases the primary motive force is multiplied by some mechanical device.

An understanding of the force and speed relationships between the input and output of a system of gears is essential. The remote control means of training and elevating guns in automatic control utilizes a system of gears following the electric or hydraulic motor rather than direct drive from the motor. Many of the standby methods of manually operating gear such as guns, turrets, rudders, directors and rangefinders depend upon the mechanical advantage inherent in a set of gears. Complex machines

such as rangekeepers are composed primarily of trains of gears, mechanical converters and mechanical integrators. In the mechanical integrator a disk is driven at a constant angular velocity ω, so that a roller of radius r_1 in contact with the disk at a distance r from the latter's axle will rotate at a rate ω_1. Thus if the constant angular velocity ω of the disk is used to represent increments of time, and the distance r is varied to represent a varying range rate, the number of rotations made by the roller shaft will be a measure of the total increment in range.

Blocks and tackle are used in hoisting boats and other heavy equipment by means of boat cranes and boat booms. It is necessary to be able to calculate not only their mechanical advantage but as well the tensions set up in the hoisting cable. In making such calculations consideration must be given to frictional forces and the inertial forces that arise in sudden starting or stopping. Frequently, the mechanical advantage afforded by a system of pulleys is used to improvise in unusual situations such as the removal of damaged gear, the taking of a strain to prevent a further failure of weakened structures, or the righting of capsized ships such as the U. S. S. Oklahoma.

Ability to make power and energy calculations is essential. There is the necessity of understanding mechanical means of measuring the output horsepower of engines, by various types of band brakes. Familiarity with the use of the equation $P = Fv$ is essential in order to study the power required to drive ships through the water and airplanes through the air. Problems of this type need not be concerned with the complex dependence of the resistance forces on surface conditions, area of the surface, viscosity and velocity. Instead, the force for any particular velocity may be obtained from a force-velocity graph plotted from experimental data on the ship involved. Such problems offer valuable training in dealing with variable forces and in the use of graphical methods of analysis. As many of the naval engineering applications of physics involve the use of graphical representation rather than the approximation of exact formulas, it is well to prepare the beginner for this transition so that he will not look upon physics as a science overfull of impractical theories. A realization of the rapid increase in power needed with increased cruising speed is important if ships are not to consume excessive amounts of fuel and if airplanes are to return from distant combat missions.

Interior ballistics is concerned with the thermodynamics and mechanics involved in propelling a projectile out of its gun barrel. Energy considerations are most important in arriving at the final muzzle velocity of the projectile. A gunnery officer must be familiar with the principles involved in the conversion of chemical energy of the explosive powder into kinetic energy of the projectile. As accurate gunnery depends upon accurate knowledge of the muzzle velocity of the projectile, he must understand the effects of temperature, projectile weight, gun wear and powder charge on this velocity. In obtaining a particular velocity it is important that the maximum pressure within the gun barrel does not exceed the rated limiting pressure. Of the energy in the powder some 45 percent is lost in heat and motion of the gases following the projectile out of the gun muzzle. Some 20 percent is transferred as heat, including friction, to the gun.

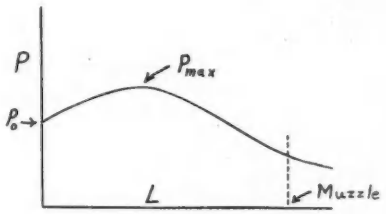


FIG. 7.

The remainder is available for conversion into translation and rotation of the projectile and translation of the gun.

The work done by the expanding gases may be obtained from the usual pressure-volume curve. In actual practice the powder burns at a rate such as to give the pressure curve shown in Fig. 7. There P represents the pressure of the gases in the gun barrel and L represents the position, or "travel," of the projectile along the gun barrel. The pressure builds up to some 3 to 5 ton/in.^{2} before the projectile starts to move. The powder continues to burn and builds up the pressure until the value P_{max} is reached. The gun must be *built up* or *radially expanded* so as to withstand this pressure. Considerations of Hooke's law and elastic limits are involved in both of these designs. The work done on the projectile is of course proportional to the area included under this curve out to the point where the projectile leaves the gun. Interesting physical

problems can be designed around either the energy available to the projectile or the mean pressure acting upon it.

Example 3. A 200-ton 12-in. 50-caliber gun fires a 1-ton projectile of radius of gyration 4 in. with a muzzle velocity of 3000 ft/sec. The rifling of the barrel is of such pitch as to cause the projectile to make 1 rev while traversing the length of the barrel.

Simple calculations show that the translational kinetic energy of the projectile is 280×10^6 ft lb, and that its rotational kinetic energy is 0.49×10^6 ft lb. Thus only an insignificant fraction of the shell's energy is rotational. From the principle of conservation of momentum, the velocity of recoil of the gun is 15 ft/sec, and hence the translational kinetic energy of the gun is 1.4×10^6 ft lb. The total energy of 281.9×10^6 ft lb can be shown to require a mean effective pressure of 25 ton/in.^{2}.

Small as the energy of recoil of the gun may appear when compared with that carried away by the shell, it is still quite large in an absolute sense and provision must be made for its absorption. This is done by means of the recoil and counter-recoil mechanism. The major portion of this recoil energy is absorbed by hydraulic brakes. As the gun recoils a piston moves through a cylinder filled with fluid. The liquid is forced through an orifice from one side of the piston to the other, thus absorbing energy and rising in temperature. A small part of the energy is absorbed in the compression of a spring or air and is then used to return the gun to its original firing position. This latter is the counter-recoil. The recoil mechanism also increases the time available for the gun to transmit its momentum to the gun mount and thereby decreases the force of the trunnions on the mount.

The impulse that the projectile and gun receive in the example cited is easily seen to be 186,000 lb sec. If this impulse were absorbed by the gun mount during the 0.033 sec that the projectile is in the gun barrel, the average force of the trunnions on the gun mount would be 5,600,000 lb. Since the average recoil distance of a gun is 3 to 6 calibers, the time available for the gun mount to absorb the impulse is at least ten times that just assumed. Such an increase in time will reduce the force to 560,000 lb or less. This problem certainly shows that if a gunnery officer is to understand the principles underlying the operation of modern guns, a considerable knowledge of mechanics is required.

Applications of momentum and impulse relationships are becoming of increasing interest and importance in naval and air warfare because of the use of jet-propelled devices. It is important for the beginner to realize that: (1) air is not required for the expelled gases in rockets to react upon; (2) there is little or no reaction on the rack from which the rocket is fired; (3) any tube or mount used in firing a rocket projectile serves primarily as an aiming device. The majority of "secret weapons" that have become generally known to the public utilize the principle of jet propulsion. The most significant are the bazooka, rocket bombs fired from airplanes, rocket projectiles fired from small landing craft, jet propulsion robot bombs and jet propulsion airplanes. Many more such devices will undoubtedly come into use as the war progresses. Therefore, it is important that the training of naval officers include an understanding of the principles involved in such devices.

Much additional information on naval applications of mechanics may be obtained by a perusal of some of the professional textbooks used in the instruction of midshipmen at the U. S. Naval Academy.[3] A physicist reading these books will find conventions and definitions different from those common to physics textbooks. Muzzle velocities are expressed in "ft-sec." Pressures often are expressed in "pounds," which serves to enhance the beginner's confusion as to the difference between force and pressure. Newton's second law and all the relationships derived from it are based on the equation $F=(W/g)a$. He may also run into errors such as the statement that a centrifugal force acts on a ship to cause the ship to heel. However, he will in addition find that a thorough training in the fundamental principles of mechanics is indispensable to a successful career in our Navy.

THE GYROSCOPE†

The general physics course at the U. S. Naval Academy treats the gyroscope as a vital element of modern naval equipment, for it is the heart of many intricate instruments. Moreover, we

[3] The books that should be of most interest to teachers of physics in the Navy V-12 Program are: Department of Ordnance and Gunnery, *Naval ordnance* (U. S. Naval Institute, 1939); Herrmann, *Exterior ballistics* (U. S. Naval Institute, 1935); Manning and Schumacher, *Principles of warship construction and damage control* (U. S. Naval Institute, 1935); Knight, *Modern seamanship* (Van Nostrand, 1941).

† By John D. Riggin, Lieutenant Commander, USNR.

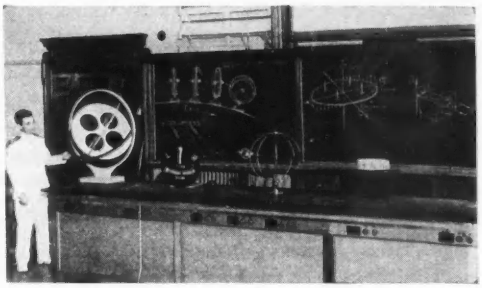
live on a gyroscope, and our weather is governed to some extent by the gyroscopic action of large air masses.

A thorough understanding of gyroscopic principles is essential to the midshipmen in several courses given at the Naval Academy. In the Department of Seamanship and Navigation the uses of the marine gyrocompass are taught, and in the Electrical Engineering Department the associated electric circuits are studied. In the Department of Ordnance and Gunnery, when such complex mechanisms as gun directors, gun sights and the like are studied, a prior knowledge of the gyroscope is assumed. The Department of Marine Engineering treats gyro aircraft instruments and the Link "blind flying" trainer. Thus it is clearly the duty of the Naval Academy Physics Committee to present the gyroscope in such a manner that every midshipman will carry with him a lasting impression of gyroscope mechanics and applications.

Inasmuch as the ultimate purpose of a fleet is to deliver explosive projectiles to the enemy as a target, any apparatus directly connected with this operation can undoubtedly be labeled vital. This is true of the gyroscope that directs the aiming of the guns in the last and toughest portion of the "delivery of explosives." In fighting off an attack by enemy aircraft, the gyroscope plays an entirely different but nonetheless effective role. In a torpedo attack against the enemy, gyroscopes help guide the aircraft to the torpedo release point, from which place the gyro in the torpedo takes over—steering the torpedo on a collision course with the enemy unit. Thus the gyroscope participates in almost every phase of gunnery.

Precise navigation of ships, insuring quick transit of the oceans and exactness in finding a rendezvous, is of hardly less importance than good gunnery. The gyrocompass, with its many repeaters, and the marine gyro automatic pilot contribute much to successful navigation.

The apparent violations of the "normal" laws of motion by the gyroscope have led to many false notions and superstitions among uninformed people. Treatments of the gyroscope in many textbooks and instruction and service manuals make somewhat erroneous statements, often ascribing mysterious properties to the gyroscope and making an adequate understanding of gyroscopic principles and effects difficult if not impossible for the reader to grasp. In teaching the midshipmen, great care is exercised to prevent



Official U. S. Navy photograph

Lecture on gyroscopes.

improper use of such familiar but erroneous gyroscopic "maxims" as:

> A gyroscope will maintain its axis fixed in space!
> A gyroscope tends to maintain its plane of rotation!
> A gyroscope, free to rotate about any axis, will remain fixed in space regardless of the movement of the earth or of the gyroscopic supporting frame!

If it is understood that a gyroscope opposes a change in position, no great harm is done. But if it is supposed that a gyroscope will maintain its axis in space, come what may, further study of the device is fruitless.

Precession offers no great difficulties. A simple vector addition of angular momentum (along the axis of spin) and angular impulse (along the axis of the applied torque) leads to the expression $T\,dt = I\omega\,d\phi$, from which it follows that $T = I\omega\Omega$. Here T is the applied torque, I is the moment of inertia of the rotor about the spin axis (a principal axis), ω is the spin velocity, $d\phi$ is the infinitesimal angle through which the spin axis moves in a plane at right angles to both the plane of the spin and the plane of the applied torque, and Ω is the angular velocity of precession. It is further shown that, in accordance with Newton's laws of motion, when a torque is applied to a gyroscope and precession occurs, a reaction couple, called the *gyroscopic couple*, is set up by the gyro which is equal in magnitude to the applied torque but opposite to it in direction. The gyroscope will suffer no deflection in the direction of the applied torque but will precess about an axis at right angles to both the spin and the applied torque. Nowhere in the fundamental equation $T = I\omega\Omega$ is there any im-

plication that the gyroscope will "remain fixed in space," but quite the contrary is shown. Given a gyroscope of finite moment of inertia and finite spin velocity, any applied torque, however small, will produce a precession of the spin axis. This should dispel once and for all any notion that the gyroscope will "maintain" its axis of rotation.

While, in the absence of any applied torque, the gyroscope will certainly maintain its axis fixed in space, so will any other object in the universe, spinning or not. The property of a body to maintain an axis fixed in space in the absence of an applied torque is not an exclusive property of the gyroscope but belongs to all masses. However, a spinning mass will be deflected much more slowly by a given applied torque than will the same mass not spinning. From this fact arise the erroneous notions regarding gyroscopic "stability."

Confusion can stem from the so-called "forced precession" problems. Here is a typical problem, reduced to its essentials.

Example. A ship rolls to the right at the rate Ω *(rad/sec). An electric generator with rotor of moment of inertia* I *(slug ft*^{2}*) and spin velocity* ω *(rad/sec) is mounted with the spin axis across the ship. The spin is clockwise when viewed from the left-hand side. Find the gyroscopic couple set up by the rotor against the bearings of the generator.*

It takes no great amount of insight to decide that the magnitude T of the gyroscopic couple is $I\omega\Omega$ (lb ft), but to find the direction is another matter. In the usual rule of spin-chases-torque or the right-hand rule of gyroscopic precession, it must be noted that the torque mentioned is the *applied torque* and *not* the gyroscopic couple. However, the latter is equal in magnitude and opposite in direction to the former. The obvious manner of attack seems to be as follows. The spin vector is directed horizontally to the right, and the precession vector is directed horizontally forward. Applying the right-hand rule in which precession, spin and applied torque are directed along the x, y and z axes of a right-handed coordinate system, one sees that the applied torque vector is directed vertically downward. The gyroscopic couple is, perforce, directed opposite to this, or vertically upward. Thus the generator shaft presses aft on the left-hand bearing and forward on the right-hand bearing.

In explaining the action of "forced precession," a step-by-step process is used. The ship of the preceding problem starts its roll to the right. A downward force is applied to the right-hand end of the generator shaft and an upward force to the left-hand end. This applied torque causes the rotor to precess about a vertical axis counterclockwise as viewed from above. This precession is stopped by the generator bearings almost as soon as it starts. However, the bearings must exert a couple to stop the precession. This couple is in the same plane (but opposite direction) as the precession and causes a new precession in the plane of the ship's roll. It is the gyroscopic reaction to this couple that is found to be equal to $I\omega\Omega$.

In considering the question of gyroscopic stability it must be remembered that a rotating mass without freedom to precess will offer no gyroscopic stability whatever. For example, the rotating machinery aboard a ship in no sense affects the ship by its effort to "maintain its plane of rotation." Only when the rotating mass is allowed to precess is the gyroscopic couple set up. Sir Henry Bessemer constructed a ship with a large cabin set on fore-and-aft trunnions. Mounted rigidly in this cabin was a large rotating flywheel which was expected to prevent the cabin from rolling with the ship. Needless to say, the cabin rolled the same as if no flywheel were present. Had Sir Henry mounted the wheel in such a manner as to allow a precession about an axis at right angles to the ship's roll axis, his device would have stabilized the cabin and he would have anticipated the Schlick (and perhaps the Sperry) ship stabilizer.

In many cases where gyroscopic stability is utilized—torpedo steering and the aircraft direction gyro, for example—the spin axis of the gyro will precess slightly under the influence of frictional torques in the vertical gimbal bearings. This very precession sets up a gyroscopic couple which opposes the friction, and the vertical gimbal will exhibit a remarkable degree of azimuthal stability. Continual application of such friction will precess the spin axis into coincidence with the torque axis and the desired stability vanishes.

A small demonstration gyroscope is used by each instructor in the Naval Academy recitations to demonstrate the afore-mentioned effects. Qualitative effects of applied

torques are easily shown, and the stability of the outer gimbal as well as the application in the various gyroscopic devices are exhibited. Each midshipman is permitted to apply torques to the gimbals and feel the surprising gyroscopic reaction.

A large model of the demonstration gyroscope is also shown, and attention is directed to the meaning of many of the terms used in discussing angular phenomena. Precession and stability are shown clearly. A somewhat smaller electrically driven gyroscope is allowed to precess under the action of a measured torque, and the rate of precession is observed. The spin velocity is measured stroboscopically and, the moment of inertia of the rotor being known, the formula $T=I\omega\Omega$ is checked experimentally.

Working models of several aircraft instruments, approximately five times actual size, are demonstrated. These models are driven by compressed air as are their prototypes. The Turn and Bank Gyro, the Gyro Directional Indicator and the Artificial Horizon comprise this group of instruments.

The marine gyrocompass is demonstrated with a small gyroscope mounted on a large rotatable globe. The force of gravity of the compass bail weight is simulated with a stretched rubber band attached to the inner gimbal and to the center of the large globe. Damping is accomplished by the eccentric connection of the rubber band to the gyro gimbal. In explaining the compass action, use is made of rather elaborate drawings on the lecture hall blackboard.

In the ships of the Navy, gyro instruments meet an important need. Foremost of these is the marine gyrocompass. The magnetic compass was never a paragon of accuracy and is not capable of meeting the requirements of modern navigation, since its motivating agency, the earth's magnetic field, is distorted and diminished by the steel of the ship. It must be supplemented by a compass less subject to error and capable of operating repeaters in various parts of the ship. The gyrocompass is the obvious answer to these requirements. Basically, it consists of a gyroscope mounted in gimbals, an almost frictionless vertical bearing, a gravitational torque system and a damping device.

The gyroscope must be given a directive force sufficient to cause its spin axis to seek the meridian and to return quickly to the meridian if deflected. The complete instrument must be essentially free from errors caused by the motion of the ship and accelerations produced by changes of course and speed. Consider first the motion of a free gyroscope placed on the earth's equator with the spin axis horizontal and the spin vector pointing east. As the earth rotates toward the east the axis of the gyroscope will appear to tilt upward. After 6 hr it assumes a vertical position. Examination of this phenomenon will show that, although the gyro axis changed its direction with respect to the earth, it remained sensibly fixed in direction with respect to the stars while the horizontal tangent plane of the earth tilted 90°. At the end of 24 hr the gyro will resume its original position. This is an illustration of the use of the equation $T=I\omega\Omega$; that is, if a very small frictional torque is applied to a rotor having a large angular momentum, the spin axis of the rotor will be deflected only a barely perceptible amount in a matter of hours.

To give the gyro a directive force, a small mass, called a *bail weight* (it resembles the inverted bail or handle of an ordinary water pail) is hung on the gimbal bearings and loosely fastened to the bottom of the gyro rotor housing. A righting torque is introduced when this mass is not directly under the center of gravity of the rotor assembly. Assume that the spin vector of the gyro is pointing east and is slightly above the horizon. The torque produced by the bail, which tends to return the spin axis to the horizontal, actually produces a precession of the spin axis from east toward north, through north and on toward west. The continued rotation of the earth brings the local vertical into alinement with the bail weight and the center of gravity of the rotor, and precession ceases momentarily. Further rotation of the earth reverses the process, and the spin axis precesses back through north and toward east. The spin axis sweeps out an elliptical cone with north as its axis. The time required for a complete oscillation about north is approximately 84 min. It can be shown that this particular period will result in the elimination of the "ballistic deflection error."[4]

In order to make the gyro axis seek and remain in the meridian, the oscillations must be damped in such a manner that the spin axis will be horizontal when in the meridian and thereby come to rest. In the older models of the Sperry gyrocompass this is accomplished by attaching the bail weight to the rotor case eccentrically (slightly to the east of the vertical axis) so that when the north end of the axis is above the horizon, a torque is introduced which will cause

[4] Ferry, *Applied gyrodynamics* (Wiley, 1933), p. 186.

that end to precess downward and, when below the horizon, will cause it to precess upward. This precession, combined with the one previously described, causes the north end of the spin axis to describe a converging elliptical spiral. After about two cycles the axis, while horizontal, reaches the meridian and thus no further displacing torque is present.

Normal rolling of a ship would set the gyrocompass assembly swinging and further complicate matters. To prevent this, some means of stabilization must be employed. In the Sperry compass, mercury ballistics are used instead of the pendulous bail weight. The center of gravity of the ballistic assembly coincides with that of the rotor and the former is thus nonpendulous with respect to the latter. In the Arma compass, two gyro elements are mounted with their spin axes approximately 120° apart. These gyros are allowed to precess about vertical axes and serve to stabilize the sensitive element.

The master gyrocompass is usually located in a special room low in the ship. Repeating elements reproduce the heading as determined by the compass in various parts of the ship, such as the navigating bridge, plotting rooms and after steering position. Many of the navigational aids, such as the dead reckoning tracer, employ gyrocompass repeaters, as does the underwater sound gear.

Gunnery data, obtained from spotting stations and fed into the various fire-control directors and computing machines, must be of the highest accuracy practicable. The bearing of the target is measured from a base line fixed relative to the earth. This base line is obtained from a gyrocompass of the most precise construction possible. It follows that while most gyrocompasses may be used for navigational purposes, only the most precise are suitable for fire-control purposes.

Elevation angles are measured from a true horizontal plane established from a gyro stable element vertical. This element, while similar to the aircraft gyro horizon, is also constructed to the highest degree of precision known to the art.[5] The sighting telescope may have its optical elements stabilized by a gyro pendulum so that the image of the target will remain stationary in the field of view irrespective of the rolling and pitching of the ship.

The automobile torpedo, operated by its own power and capable of underwater speeds as high as 40 knots at ranges up to 8000 yd, must be steered on a collision course with the target. This course must be maintained regardless of how the torpedo is deflected by waves. Steering is accomplished by means of a free gyroscope which, in the absence of appreciable torques, will maintain its spin axis direction approximately fixed in space. Before the torpedo is launched the gyro is locked with its spin axis in line with the fore-and-aft torpedo axis. In the fraction of a second after the torpedo is fired and before it leaves the launching tube, the gyro is accelerated to a speed of about 10,000 rev/min and then unlocked. A jet of compressed air directed against small buckets cut in the gyro rim keeps it spinning at the required speed.

If, during a normal straight run, the torpedo is deflected from the heading it had when the gyro was unlocked, the horizontal angle between the gyro axis and the torpedo axis is no longer zero. This angular displacement controls the admission of high pressure air to the steering engine and returns the torpedo to its correct course. Operation of the steering-engine valves should not result in the application of an appreciable torque to the gyro, as such torques might cause the gyro to precess from its proper heading and thus steer the torpedo on an incorrect course. Small frictional torques, however, introduced about the vertical axis of the gyro owing to the yawing of the torpedo will cause the spin axis to precess, *but only in a vertical plane*. It is the orientation of this vertical plane that is of importance in the steering of the torpedo. It can be seen that a limited amount of precession in this plane will not affect the heading of the torpedo.

The effect of gyrodynamics on the drift of a spinning projectile is not fully understood,[6] but it is supposed that drift and spin are in some way connected, for reversing the direction of the projectile's spin reverses the direction of its drift. Some authorities are of the opinion that the spin of the projectile has a detrimental effect on the stability of the projectile in flight.[7]

Gyroscopic flight instruments are an important contribution to the "blind" operation of aircraft. Without the information furnished by such instruments it is impossible for a pilot surrounded by dense clouds or fog to determine the orientation of his aircraft. He has no way of

[5] Reference 4, p. 128.

[6] Gray, *Gyrostatics and rotational motion* (Macmillan, 1918), p. 148.

[7] Cordeiro *The gyroscope* (Spon, London, 1913), p. 85.

knowing whether the aircraft is inclined to the horizontal, turning or even inverted. An aircraft cannot be flown instinctively, that is, without reference to a horizon. It has been reported that even seagulls, attempting flight in a fog, become confused, stall and make "crash landings."

Blind flying is quite practical with proper gyroscopic instruments. The gyro horizon tells the pilot the attitude of the aircraft with reference to the true, although perhaps invisible, horizon. The direction indicator enables him to hold a steady course and make precision turns. Modifications of these two instruments, when connected to the control surfaces of the aircraft through a hydraulic or electric power system, actually fly the aircraft without assistance from the pilot and are known collectively as the gyro pilot. A third instrument, the gyro turn and bank indicator, shows the rate of turn of the aircraft.

The flux gate compass, a gyro-stabilized magnetic compass,[8] is rapidly replacing the conventional type of magnetic compass in aircraft. This new compass possesses a high degree of sensitivity and is especially useful in high latitudes where the horizontal component of the earth's magnetic field is comparatively weak. Turn and bank errors inherent in the conventional compass are absent in the flux gate compass.

Precision bombsights employ the gyroscope as a stable reference.[9] The fundamental reference principle is similar to that of the fire-control director used in surface ships.

Current aircraft gyro instruments are operated by alternating current[10] in contrast with earlier types, which used a slight vacuum obtained from Venturi tubes or an electric vacuum pump. The vacuum developed by the Venturi is inadequate at today's high altitudes.

As can be inferred from this discussion, the embryonic naval officer may expect to meet the gyroscope at every turn on Navy ships and aircraft. One of the aims of the Naval Academy physics course is to make every midshipman aware not only of the theory and applications of the gyroscope but also of its expanding use in our expanding Navy.

[8] News Week **22**, 97 (Oct. 25, 1943); Flying **34**, 69 (Jan. 1944).
[9] Life Mag. (Jan. 24, 1944).
[10] Sci. News Letter **45**, 105 (Feb. 12, 1944).

SOUND†

Appropriately enough, sound in the Navy has come to mean, for the most part, underwater sound. By the shift of emphasis from air to water the general nature of wave motion is made clearer to the student. Such fundamentals as vibrating sources, the velocity of compressional waves, the relation $v=f\lambda$, the laws of reflection and refraction, and interference are of course given primary consideration. Some phenomena, however, such as diffraction, which is not ordinarily treated in connection with sound in elementary textbooks, and the Doppler effect, which is usually presented mostly as a scientific curiosity, can be shown to have a very real significance.

Long before sound in the Navy "went underwater," mariners were using a knowledge of the speed of sound as an aid to navigation. Along the banks of some inland waterways large "sounding boards" are arranged to parallel the shore line. A navigator approaching such a shore in a fog can obtain a very good estimate of his distance by sounding his whistle and timing the interval until the sound returns as an echo. In some localities precipitous banks rise to a considerable height above the water and serve to reflect the sound waves.

Example. A vessel making 4 knots and approaching land in a heavy fog is being navigated by means of echoes from sounding boards along a straight shore line. The air temperature is 68°F. At 10 o'clock the echo interval is 10 sec; 15 min later it is 6 sec. How much should the course of the vessel be changed in order for it to be parallel to the shore line?

Many lightships send out characteristic warning signals by means of underwater bells or oscillators. Vessels equipped with underwater sound receivers can detect such signals at distances in excess of 50 mi. Some lightships send out radio signals and underwater sound signals simultaneously, thus making it possible to determine the distance from a receiving vessel to the lightship by measuring the time interval between the arrivals of the two signals.

† By E. R. Pinkston, Lieutenant, USNR.

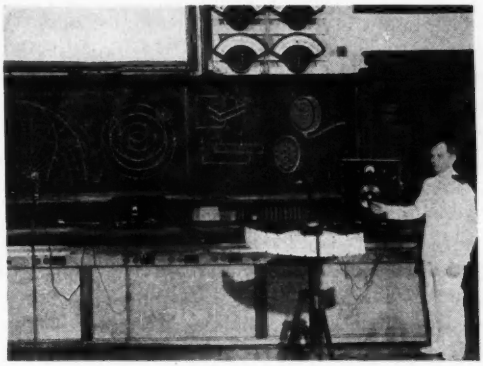
An interesting application of distance determination by synchronous signaling is employed by a beacon station in the Firth of Clyde. The station broadcasts a recording which first gives the name of the station and then a series of blasts synchronized with those of the station fog signal. This is followed by a voice counting in miles at such a rate that the distance to the receiver is given at the instant when the last blast of the fog signal is heard directly through the air.

Perhaps the most useful and practical of all peacetime applications of underwater propagation of sound is the determination of depth through a knowledge of the speed of sound in water. In the most common installation a pointer moves rapidly around a dial graduated in fathoms. As the pointer passes the zero mark a sound pulse is sent toward the bottom by a powerful underwater oscillator. When the returning sound is received, an electric circuit is set up causing the pointer to be illuminated momentarily as it passes the proper fathom mark on the dial.

Example. The dial of a sonic echo-sounding device is graduated so that 1 rev of the pointer represents a depth of 100 fathoms (600 ft). How many revolutions per second should the pointer make if the speed of sound in sea water is 4800 ft/sec?

When a device or method is employed that makes it possible to confine most of the energy from a sound source to a narrow beam, the time of travel of a sound pulse can be used to determine distances in given directions. In 1912, M. L. F. Richardson took out a British patent suggesting "apparatus for warning a ship at sea of its nearness to large objects wholly or partially under water." Since that time, and more particularly in recent years, the method of detecting and determining the distance and direction of submerged bodies by means of sound waves has been employed with considerable success. The same method appears to offer promise in the detection and location of schools of fish.

Most of the afore-mentioned applications of sound waves in distance determinations utilize not only the finite velocity of the wave motion but also its reflection. In one method of depth finding, useful mostly in shallow water, a sound receiver with directional characteristics is placed near the bow of the vessel to determine the direction of the vessel's propeller sounds after they have been reflected by the ocean bottom.[11] In this instance a knowledge of the speed of sound is not required. The laws of reflection are applied and the problem is solved by trigonometry.



Official U. S. Navy photograph

A lecture on sound.

Example. Hydrophones located near the bow of a vessel, 20 ft below the waterline and 300 ft from the propellers, receive the sound of the propellers at maximum intensity along a direction downward and toward the stern that makes an angle of 60° with the vertical. Calculate the depth of the water, assuming the ocean floor to be horizontal.

It is interesting to note that there are three paths along which the sound wave may travel from the propeller to the receiver: (1) by reflection from the bottom; (2) directly, alongside the vessel; and (3) by reflection at the surface of the water. The intensity of the sound received over the longest path (from the bottom) is considerably higher than that over either of the other paths. This is because the sound wave undergoes a change in phase upon reflection at the surface and thus interferes destructively with the sound traveling the direct path.

Example. An underwater oscillator of frequency 480 vib/sec is located near the stern of a vessel, and a receiver is located near the bow, 300 ft away. Each is 10 ft below the water line. Calculate in terms of wavelength the phase difference of sound signals arriving at the receiver directly and by reflection from the surface of the water.

[11] H. C. Hayes, Proc. Am. Phil. Soc. **59**, 317 (1920); **63**, 134 (1924).

In connection with a study of the geometric laws of reflection of sound at the bounding surfaces between two different mediums it is of value to indicate in general how the energy of the wave motion is partitioned between the reflected sound and the sound that travels on into the second medium. Thus, when a sound wave impinges upon a boundary between two mediums having widely different radiation resistances, such as an air-water boundary, there is almost complete reflection of the sound energy regardless of which medium transmits the incident sound.

Refraction is usually associated with the passage of a wave from one medium into another in which the velocity of propagation is different. However, much of the refraction of sound waves that has to be dealt with in practical applications occurs within a single substance in which the changes in properties are more or less gradual. It has long been recognized that in a heavy fog the direction of a sound source cannot accurately be deduced from its apparent direction at the receiver. It is not unusual for a lookout to hear a fog signal on the starboard quarter and find when the fog lifts that the source of the signal is actually on the port quarter.

The principal factors affecting the speed of sound in sea water are the temperature, the salinity and the pressure. As a general rule, temperature changes account for most of the refraction of an underwater sound signal. In certain localities there is likely to exist a warm layer of water at some distance below the surface. Such a temperature inversion causes the sound to be reflected back and forth between the surface and the warm layer, the latter being more or less impervious to the sound signal.

The phenomena associated with diffraction have played an important part in the development of the directional transmission and reception of sound. Those who would understand the practical dependence of beam transmission upon high frequency sound need to know something about diffraction effects. It is important for the student to understand something of the relationship between the physical dimensions involved and the wavelength of the sound even if he is not prepared to follow rigorous mathematical treatment of the problems. Most of the sound energy radiating from a "piston" source is confined within a primary beam of semi-angle $\sin^{-1} 1.22\lambda/D$, where λ is the wavelength in the transmitting medium and D is the diameter of the source. When the diameter of the source is less than $\frac{1}{4}\lambda$ the source can be regarded as a "point," radiating spherical waves; but if its diameter is in the neighborhood of 10λ, most of the sound is confined within a cone of comparatively small angle. The following problems illustrate one of the difficulties in obtaining sound sources that are directional.

Example 1. The range of sound transmission in air is a maximum for a frequency of about 200 vib/sec. What should be the diameter of a "piston" source that will produce a "useful beam width" of 18° in air of temperature 6°C?

Example 2. It is desired to radiate sound from an underwater source so that most of the energy is concentrated within a cone of angle 20°. Compare the necessary diameters of two such piston sources whose frequencies are 2000 and 20,000 vib/sec, respectively. Assume the speed of sound in the water to be 4800 ft/sec.

There are many methods of demonstrating the relation between the frequency and the directivity of a given source. In a lecture demonstration used at the Naval Academy an ordinary loudspeaker driven by a beat-frequency oscillator is placed on a turntable having a dial and pointer sufficiently large for the audience to see. A few feet away on the axis of the source is placed a stationary microphone connected through an amplifier and a rectifier to a large wall meter. Each student plots on a sheet of polar coordinate paper the meter readings *versus* the angular position of the microphone with respect to the axis of the speaker, for several different frequencies. Although the quantitative results of such an experiment are not emphasized, the students do become aware, some of them for the first time, that, for a given source, sounds of high frequency are more directional than those of low frequency. This experiment serves to emphasize one of the reasons why sounds of high frequency have come to be used in underwater signaling.

When it is desired to render audible the high frequency signals generated by a supersonic transmitter, recourse is had to difference tones. Suppose, for instance, that soundings are being made by the echo method and that sound pulses

of frequency 50,000 vib/sec are traveling to the bottom and back. It is desirable to hear the echo as well as to see the flashing pointer as it passes the proper fathom mark on the dial. In order to accomplish this the returning pulse of 50,000 vib/sec can be combined in the receiver with another signal of frequency 49,000 vib/sec. This combination results in an audible difference tone of frequency 1000 vib/sec.

J. O. Perrine,[12] in a recent discussion of the Doppler effect, describes the case in which a reflecting surface moves with a speed of 50 ft/sec toward a receiver located near a fixed source having a frequency of 500 vib/sec. The resulting frequency is 546 vib/sec. When the source in this case emits a pulse of 50,000 vib/sec, the frequency of the reflected pulse is 54,600 vib/sec. If both the outgoing pulse and the echo are rendered audible by combining them with a frequency of 49,000 vib/sec, the difference tone has a frequency of 1000 vib/sec for the outgoing pulse and 5600 vib/sec for the echo. Such a magnification of the Doppler effect indicates that even a moderate velocity on the part of the reflector will produce considerable change in pitch. In a lecture demonstration of this effect two sources of fairly high frequency are adjusted so that the frequency difference is approximately 600 vib/sec. The sound from one of the sources is received directly by a microphone while that from the other source arrives at the microphone after reflection from a plane surface. The signal received by the microphone is rectified and rendered audible by amplification. With the reflector stationary the combination of the two high frequencies produces a difference tone of a considerably lower pitch. Motion of the reflector is accompanied by a marked change in the pitch of the difference tone.

The directional property of the early underwater listening devices was attained by means of the binaural effect.[13] One of the earliest of such devices had two hollow, spherical, rubber receivers fixed at the horizontal ends of a T-shaped frame. The frame could be rotated about its vertical axis, and the receivers were connected by two tubes to stethoscopes. With the horizontal tube trained athwartship an azimuth circle indicated a relative bearing of zero. On hearing a sound which appeared to be on his starboard hand, an operator wearing the stethoscopes would rotate the device to starboard until the sound was centered. In practice the ship always had to be stopped, as at speed of 2 knots or more the water resistance made it extremely difficult to rotate the apparatus and there was a considerable amount of water noise. Later devices of a similar nature had as many as 16 receivers, eight connected to each ear.

A great improvement was the addition of a binaural compensator by means of which the effective paths to the two ears could be equalized without having to rotate the receivers. Thus with the receivers mounted permanently on the hull of the ship a sound pulse could be made to arrive at both ears simultaneously by increasing one of the air paths to compensate for the difference in the paths in the water. The compensator could thus be graduated in degrees of bearing.

Example. The two ears of an acoustic listening device are 6 ft apart athwartship. If sound is received from a relative bearing of 130°, how much compensation would have to be provided, and for which ear, when the air temperature is 25°C and the speed of sound in the water is 4800 ft/sec.

In electric devices that utilize the binaural effect the compensation is accomplished by shifting the phase relation of the currents in the two circuits to balance the phase difference of the sound waves.

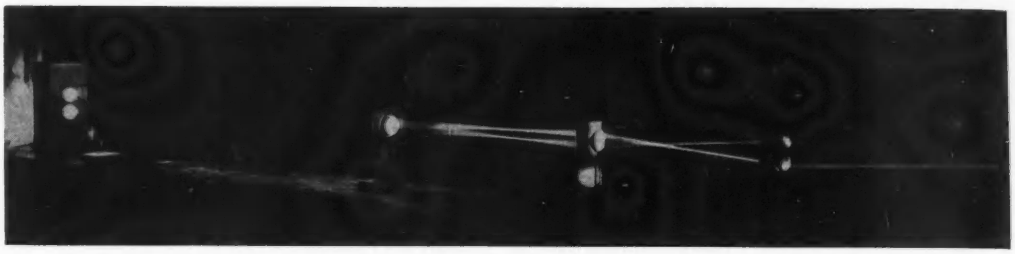
Obviously the naval officer of today needs to have a good understanding of the fundamental principles of sound. Because the midshipman has little opportunity to study the subject after he finishes elementary physics, we consider it suitable wherever possible to use naval applications to illustrate the principles being studied. Sound principles should suffer no ill effects from immersion in water.

OPTICS†

One of the rules of nature is that the new replaces the old. It is, however, quite premature to say that optical devices have been replaced by the more recent and spectacular inventions

[12] J. O. Perrine, Am. J. Phys. **12**, 23 (1944).
[13] H. C. Hayes, Proc. Am. Phil. Soc. **59**, 1 (1920).

† By Ralph A. Goodwin, Lieutenant (j.g.), USNR.



Official U. S. Navy photograph

A lecture-experiment in light.

of this war. There are many jobs that must still be done with the aid of optical devices and several more in which the optimum accuracy of optical methods is still unrivaled. Rangefinders, binoculars and periscopes are still manned at sea.

Of all the optical systems, that of the human eye is, of course, the most important to a naval officer. Not only must an officer have very good eyesight, but he must be able to use his eyes at their greatest efficiency. It is not sufficient simply to be able to make out objects at all distances when visibility is good. The men on board a man-o-war in enemy waters *must* be able to see better than the enemy by day and by night. The failure of one man to use his eyes at their peak of performance may jeopardize the safety of the ship.

A course in general physics cannot allot much time to the study of the eye and the process of vision, but there are several facts of particular importance that should be known to every naval officer. First, two facts about the dark-adapted eye should be remembered: (1) considerable time must elapse before the eye reaches its maximum sensitivity when the illumination is changed from normal to low levels (approximately 1 lu/ft^2); (2) as the eye becomes dark-adapted, the maximum of the sensitivity curve shifts approximately 500A toward the blue end of the spectrum, the long wavelength cut-off shifting toward the blue and the blue end of the curve remaining fixed.

The first of these facts is well known and fully appreciated by anyone who has entered a motion picture theater on a bright afternoon. It might spell disaster, however, to a naval officer who, roused from sleep in the middle of the night by the sounding of general quarters, turns on a light to find his cap.

Several consequences of the second fact—the Purkinje effect—are immediately evident. Any light that must be shown where it could be seen by the enemy should obviously be a deep red rather than a blue. The shift of the long wavelength cut-off also provides a means whereby dark adaptation may be maintained at higher levels of illumination. Naval pilots are supplied with red goggles while studying their maps and charts, thus leaving their eyes still dark-adapted when they go to their planes for a take-off.[14]

Implied also in this Purkinje effect is the whole subject of rod and cone vision. Since, at low illumination levels, the fovea is practically insensitive, whereas the extrafoveal regions are highly sensitive, the visual process at low illumination differs from that for normal illumination. A sailor standing the lookout watch at night must overcome the daytime habit of focusing on the fovea, and he learns that movement of the image on the retina increases his retinal sensitivity. The task of standing the midnight watch on a battleship cruising in enemy waters requires training and some knowledge of physics.

One simple daytime use of the eye can be tried by anyone. While standing or walking outdoors try to estimate the relative bearing of various objects to within 5°. It will be found that it takes a little practice.

One of the difficult jobs in teaching any part of elementary physics is to get the student to apply himself seriously to the requisite "spade work" that must be thoroughly done before the

[14] D. H. Jacobs, *Fundamentals of optical engineering* (McGraw-Hill, 1943), p. 83.

more interesting topics and practical applications can be studied. At the Naval Academy the study of plane mirrors and reflecting prisms is enlivened by using ships as objects—a reverted or inverted image of a ship being probably more impressive than an arrow similarly situated. The sextant is an obvious application of the rotating mirror problem—and it is always mentioned that Michelson, a graduate of the Academy, was an instructor here when he first used a rotating mirror for determining the velocity of light.

Interest in the discussion of the use of nonreflecting films is always enhanced if the amount of light lost by reflection in a periscope system is calculated roughly. There is also a still more practical reason for mentioning this phenomenon to future naval officers. Everyone can imagine the steps that an uninformed seaman (or officer) might take to polish the lenses of his binoculars if they look a little off color.

It has never been found possible to devote much time to the study of polarized light, but the case of polarization by reflection is of vital importance to men at sea. The amount of light reflected from the surface of the ocean is always considerable, and it is particularly annoying when, in a glaring expanse of ocean, one is looking for a small object, such as a periscope. Polarizing filters are the obvious solution to the problem. It is interesting to imagine oneself equipped with a pair of binoculars and polaroid filters and stationed at the observation window of a baby blimp on submarine patrol. At what angle from the vertical would one have the best chance of sighting a submerged submarine? Over how large an angle would the filter be useful? Now imagine yourself on the bridge of a destroyer and search in a similar way for aircraft or surface vessels.

Nonpolarizing filters also find some use. The exact type will depend on the situation, but they could be used for detecting camouflage, looking at bright objects such as searchlights, and protection of the eye from infra-red and ultraviolet radiation.

Telescopic systems have a great many uses aboard ship. Inasmuch as most service optical instruments are designed for taking measurements, considerable attention should be given to the formation of real images, the placement of reticles, magnifying power and proper focusing. One important use of the telescope is as a gun sight. The textbook on naval ordnance used at the Academy lists three advantages of the telescopic sight: (i) the eye is focused for only one distance, instead of successively accommodating for more than one, as with the open or peep sight, and when the telescope is properly adjusted, small displacements of the eye from the sighting axis will not affect the accuracy of pointing; (ii) for a specified accuracy, the use of the telescopic sight will increase the range by a factor equal to the magnifying power of the gunsight, but limiting the useful magnification by the smallest field of view that can be used in practice; (iii) the size of the exit pupil does not affect the accuracy of sighting as does the enlargement of the hole in a peep sight, and with a telescopic sight the magnifying power and size of objective can be chosen in such a way as always to utilize the full area of the dilated pupil.

The straight telescope is not the only type in use. Since it is undesirable to lessen the protective qualities of a turret or gun shield by cutting a slot through it to permit use of a straight telescope, the sighting is done with prismatic telescopes—periscopes. Antiaircraft guns can be equipped with a prismatic eyepiece which allows the gunner to assume a more comfortable position.

The task facing a gun crew when it is called upon to hit a specified target is a good example of the operational use to which optics is put. Dual telescopic sights are used on all guns except small arms. One telescope is mounted on each side of the gun, and the two sights are rigidly connected by a device called a sight-yoke. The gun is set at zero elevation, and the two telescopic sights are adjusted so that their lines of sight intersect the axis through the bore of the gun at a definite distance. This is accomplished by mounting a third telescope in the breech of the gun and sighting it along the geometric axis of the gun. *Naval Ordnance* describes 12 steps which must be taken to accomplish this "bore-sighting." When the range of a target is given, the sight-setter moves the sight-yoke so that the two telescopes will be pointing at the target when the gun is at its correct elevation. Corrections for the deflection of the projectile at right angles to the line of fire are made by moving the sight-yoke to the left or right. One telescope is manned by the "trainer," who changes the position of the gun in azimuth and keeps his vertical cross-hair on the target. The second telescope is manned by the "pointer," who keeps his horizontal cross-hair on the target by changing the elevation of the gun.

The binocular telescope is probably the most frequently used optical system. Certain officers on board ship are using binoculars almost continuously during the hours they are on duty. It is important that they know how to focus the instrument so that continued use will not produce severe eyestrain. This point is an excellent argument in the discussion of focusing a telescope for parallel light. It may be remarked that the approved procedure is to start with the system adjusted so that the correct focus can be reached without the image passing through the so-called distance of distinct vision (25 cm), for it is this image that the inexperienced observer may automatically bring into focus. Naval officers should also understand the meaning of the system used for rating binoculars, such as 8×30, and so forth; they should appreciate the stereoscopic advantage of the instrument, and at least be fully impressed with the fact that Navy binoculars are precision instruments.

A discussion of the naval uses of optics would not be complete without mention of the rangefinder and the periscope. While a first course in physics does not usually take up optical systems as complex as that of the rangefinder, we have had rather good luck in doing so, probably because our students have an opportunity to see and operate the instruments in another department. A comparison of the relative advantages of the coincidence and stereoscopic rangefinders will vary somewhat, depending on one's source of information. The starting point of the argument is always the Battle of Jutland, which was fought with 9-ft Barr and Stroud coincidence rangefinders on the one side and 3-m Zeiss stereoscopic rangefinders on the other. This information has come from British and American sources, and the conclusions arrived at by these sources are in general agreement. Both instruments are capable of a precision of about 12″ of arc; the stereoscopic instrument is superior when ranging on indistinct objects or upon rapidly moving objects or irregular shapes. The weakness of the stereoscopic rangefinder as a service instrument apparently arises from the fact that emotional strain during battle adversely affects the precision of the rangetakers.

Periscopes have been designed with built-in rangefinders, either vertical or horizontal, with fixed and with rotating prism systems, with prisms that enable the conning officer to search for surface craft and aircraft, and with a prism system for scanning the entire horizon at once.[15] In addition to the optical systems necessary to bring an object into view, a periscope must be equipped with apparatus for determining the range and bearing of the object. Generally speaking, our submarines do not employ the rangefinder principle for getting the range but rely upon the angular measurement of the height or length of the object and the submarine commander's knowledge of the dimensions of the target.

Because of lack of space many naval applications have not been mentioned—even such important devices as searchlights and signaling lights, and such important subjects as photography and illumination. It is hoped, however, that teachers of physics may again have been assured that physics is physics, and that once the fundamental principles are mastered the application of these principles in practice can proceed.

[15] Glazebrook's *Dictionary of applied physics* devotes about 25 pages to a description of the various optical systems that are employed in periscope design.

THE science teacher should always teach that science has made two fundamental contributions to modern life: it has given man the choice between want and abundance; and it has freed him from irrational fear. Man has not quite conquered fear, but through science he has freed himself from the tyranny of ancient superstition and is gradually coming to understand his own inner fears. The scientific conception of the nature of the world and of man can free man's mind just as the scientific control of matter and energy has freed his hands.—ROBERT J. HAVIGHURST, *Sch. Sci. and Math.* **44**, 120 (1944).

Reproductions of Prints, Drawings and Paintings of Interest in the History of Physics*

19. Portraits of William Gilbert (1544–1603)

E. C. Watson
California Institute of Technology, Pasadena, California

THE 400th anniversary of the birth of William Gilbert of Colchester, a man who—as Priestley so quaintly wrote in 1767—

Plate 1. Panel portrait of William Gilbert. (From the panel portrait in the possession of the late Silvanus P. Thompson.)

"may justly be called the father of modern electricity, though it be true that he left his child in its very infancy," occurs this year.[1] It is therefore fitting to reproduce his likeness now in commemoration of this anniversary.

The only contemporary portrait of Gilbert known to exist at the present time is a small panel painting discovered by the late Silvanus P. Thompson. It was used by Charles Singer to illustrate an article entitled "Dr. William Gilbert (1544–1603)" which was published in the *Journal of the Royal Naval Medical Service* for October, 1916. Plate 1 is a very poor reproduction made from a reprint of this article. A better reproduction will appear later in this series if one can be obtained.

An original portrait, probably painted by Cornelius Jansen and bearing the date "1591, aetatis 48," is mentioned by Hearn in his

Plate 2. Engraved portrait of William Gilbert. (From the engraving by Clamp published by S. & E. Harding, May 1, 1796.)

* The first article in this series appeared in volume 6 (1938), page 112; the eighteenth article, in volume 9 (1941), page 307.

[1] The date of Gilbert's birth is usually given as 1540, since the mural tablet placed by his brothers over his burial place in the chancel of the church of Holy Trinity, Colchester, states that he died in 1603 in the 63rd year of his age. However, Silvanus P. Thompson, who is the chief authority on the life of Gilbert, considered the correct date to be May 24, 1544.

PLATE 3. GILBERT SHOWING HIS EXPERIMENTS ON ELECTRICITY TO QUEEN ELIZABETH AND HER COURT. (From the painting by A. Achland Hunt, Town Hall, Colchester, England.)

Letter containing an Account of Some Antiquities between Windsor and Oxford, with a list of the several Pictures in the School Gallery Adjoining the Bodleian Library (1708), p. 33.[2] This is the painting which GILBERT is said to have ordered made of himself for presentation to the University of Oxford. A manuscript entry at Oxford, however, states that it was removed as decayed in 1796. There remains only a poor engraving by CLAMP, made in 1796, and not true to the original portrait in several details.[3] However, this engraving, which is reproduced in Plate 2, has preserved something more of GILBERT's outward appearance than his pointed beard, ruff and high hat. "The keen straight-forward searching glance, the twinkling play of good-humoured sarcasm, ready to vent itself on all 'old wives' gossip' and 'foolish vanities,' the frank, fearless, open countenance, intolerant only of shams and frauds—all these characteristic traits of the man are not untraceable in the portrait."

ARTHUR ACHLAND HUNT made use of CLAMP's engraving for his well-known historical painting of GILBERT showing his experiments on electricity to Queen Elizabeth and her court, which is here reproduced in Plate 3. This painting was presented by the Institution of Electrical Engineers to the Corporation of Colchester on December 10, 1903, the 300th anniversary of GILBERT's death.

The word picture of GILBERT given by THOMAS FULLER in his *History of the Worthies of England* (1662) is worth quoting in this connection. The quaint and witty style is characteristic of FULLER, but he states that his information came from a near kinsman of GILBERT's. FULLER writes as follows:

> "He had the Clearness of Venice Glass without the Brittleness thereof, soon ripe and Long Lasting in his Perfections. . . . One saith of him that he was Stoicall, but not Cynicall, which I understand Reserv'd, but not Morose, never married, purposely to be more beneficiall to his Brethren. Such his loyality to the Queen that as if unwilling to survive, he dyed in the same year with her 1603. His stature was Tall, Complexion Cheerfull, an Happiness not ordinary in so hard a Student and retired a Person. . . .
>
> "Mahomet's tomb at Mecha is said strangely to hang up, attracted by some invisible Loadstone; but the Memory of this Doctor will never fall to the Ground, which his incomparable book *De Magnete* will support to Eternity."

[2] See also Poynter's *Oxoniensis Academia* (1748), entry No. 74, and A. à Wood, *History and antiquities of the University of Oxford* (1796), vol. II, p. 96.

[3] This engraving was published by S. and E. Harding in the *Biographical mirror*.

Oscar Milton Stewart, 1869–1944

PHYSICISTS who began their professional activities about the end of the last century found themselves in an intellectual atmosphere very different from that which prevails now. Three of the great American names of that time were WILLARD GIBBS of Yale, HENRY ROWLAND of Johns Hopkins and ALBERT MICHELSON of Chicago. ROWLAND died in 1901 and Gibbs in 1903. MICHELSON lived until 1931. Although the foundations of the electron theory had already been laid by J. J. THOMSON and others, and x-rays and radioactivity had been discovered, it was still a period of what we now call "classical" physics. The ether was tenaciously held to. The Michelson-Morley experiment, the character of blackbody radiation, and the nature of bright-line spectra, were unsolved mysteries. The photo-electric effect was known, but not at all understood. EINSTEIN, DE BROGLIE, SHRÖDINGER and HEISENBERG were unknown or almost unknown names. PLANCK was known and highly esteemed, but his early paper suggesting that radiation occurs by quanta was not published until 1899, and then was rather coldly received.

Men who began the study of physics as late as 1915 cannot readily appreciate what an impact was made upon the scientific philosophy of that earlier generation by the theory of relativity and quantum mechanics. The intellectual upheaval was like that suffered earlier when a devoutly religious society was confronted for the first time by the biological doctrine of evolution.

OSCAR MILTON STEWART was one of those who lived through that brilliant and fruitful, but trying, period. He was born in 1869 at Neosho, Missouri. His brother, GEORGE WALTER, professor at the University of Iowa, came a few years later. A still younger brother, VICTOR, became a successful professional photographer. It was VICTOR who is responsible for the only extant good portrait of OSCAR, which is reproduced here. VICTOR died a number of years ago. The only sister, MRS. E. L. MORGAN, is still living.

OSCAR attended DePauw University, receiving the Ph.B. degree in 1892. For graduate work, he went to Cornell University, obtained the doctor's degree in physics in 1897 and remained as instructor till 1901. It was during this period that he married MISS ESTELLE WILLIAMS, and their only child was born, a son named LAWRENCE. Much later, in 1928, DePauw honored PROFESSOR STEWART with the honorary degree of Doctor of Science.

During the nineties, the University of Missouri, at Columbia, went through a general reorganization under a new president, RICHARD H. JESSE.

OSCAR MILTON STEWART

He brought a number of promising young men to the University, including OSCAR STEWART, who was appointed assistant professor of physics, in full charge of the department, in 1901. A few years later he became full professor, and he remained as the guiding spirit of the department until his retirement in 1940.

During his 39 years at Columbia he suffered two severe buffets of fate. His son, LAWRENCE,

came within the age-brackets for military service during the first World War, and was sent to the Great Lakes Training Center; there he fell victim to the severe influenza epidemic of 1918.

MRS. STEWART's health failed noticeably during the twenties and thirties, and she died suddenly of a heart ailment, shortly before the retirement of her husband. Much against the wishes of his friends PROFESSOR STEWART continued to live alone in the old home, and the last four years of his life were not happy. He himself began to have circulatory troubles, and he suffered a light stroke within a few months of MRS. STEWART's death. He continued to have small brain hemorrhages until his death on May 17, 1944.

PROFESSOR STEWART published some very good research articles from time to time, chiefly in the fields of electrical conduction in gases and the theory of instruments; but he will be remembered mostly as a very effective teacher. He was always anxious to encourage research in the department, but insisted that research activities should not interfere with efficient and conscientious teaching. He was exceedingly patient and sympathetic with young people. The cordiality which his students of many years ago still feel for him is remarkable. His character was somewhat reserved, with a marked distaste for emotional display, but in spite of that he was a "good mixer" in a quiet, dignified way, well known and liked by the physics fraternity all over the country, and also quite popular with the townspeople of Columbia. This last is saying a great deal for a faculty man. Visitors to the University, ranging from immature graduate students to distinguished scientific men, were always impressed by his courteous treatment of them and the trouble he would take to talk to them or show them whatever they wanted to see.

He was not at all an athletic man. The most I have ever known him to do in the way of sports was to play a very mediocre golf game and paddle a canoe. Yet he was much interested in professional or college sports, and I have often heard him discourse on some fine point in baseball. He served for years on the faculty committee for intercollegiate athletics.

He was much interested in classical and semi-classical music, though so far as I know he never sang or played an instrument. He was a successful flower gardener, and he also knew practically all the local birds by their appearance or their songs.

His college textbook of physics, now in the fourth edition, has been widely used and has been very profitable for the publishers. He spent an enormous amount of time eliminating errors and improving the presentation.

For two different three-year periods he was a member of the editorial board for *The Physical Review*, and during one period, for the *American Journal of Physics*. He was an early member of the American Physical Society, of the American Association of Physics Teachers and of the Society for the Promotion of Engineering Education, and also held membership in the honorary societies Sigma Xi, Phi Beta Kappa and Tau Beta Pi. His college social fraternity was Phi Kappa Psi.

When I first met him he impressed me as being rather liberal in his social and political views, but there is no doubt that he became more conservative as he grew older. He was intelligent, and also fortunate, in investments, so that he became rather wealthy for a college professor.

When his will was probated it was found that he had left to the University a sum which is expected to yield about $1000 a year, to be administered for the benefit of the physics department. Rigid instructions for the expenditure of this interest were not stated, but the will recommends the establishment of scholarships for undergraduate or graduate students majoring in physics, and also the bringing of lecturers to the University.

The Board of Curators have decided to name the physics building after PROFESSOR STEWART. This action is particularly appropriate because he was the person most responsible for its design. The University community certainly has many reasons, besides the sense of personal loss of a valued friend, to hold in remembrance the man who so ably headed one of its important departments.

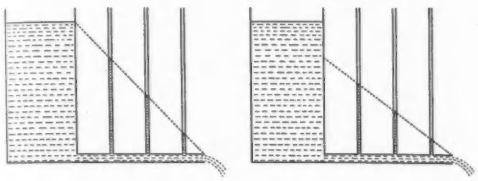
H. M. REESE

NOTES AND DISCUSSION

Loss of Head in Fluid Motion

H. W. FARWELL
Columbia University, New York, New York

POSSIBLY this is the poorest way to behave when a typographical error has been encountered in a textbook. The best seems to be to drop a kind word to the author, but if that step has been neglected, and the same error appears in other books, it is certainly inconvenient to carry Form Letter 0104*A* in the files for such purpose.



FIGS. 1 AND 2. Two diagrams appearing in different, current textbooks.

Several textbooks of general physics, all of 1943 vintage, carry a figure to illustrate the loss of head when a viscous liquid moves through a pipe. There are two figures in use, and both are reproduced herewith (Figs. 1 and 2). Of course the student should not be expected to believe that both are correct.

A Simple Form of the Clément and Désormes Apparatus

GILBERT HENRY
The University of Georgia, Athens, Georgia

AN interesting method for determining the ratio of the specific heat of a gas under constant pressure to that under constant volume was devised by Clément and Désormes in 1819.[1] Their apparatus, modified for laboratory use, consists of a large glass bulb with an airtight poppet valve and a manometer attached. A description of this apparatus can be found in almost any standard textbook on heat.[2] The ratio of the specific heats is $\gamma=\Delta h_1/(\Delta h_1-\Delta h_2)$, where Δh_1 and Δh_2 are the experimentally determined differences between the manometer column heights at the beginning and at the end of the experiment.

The apparatus described herein (Fig. 1) is extremely simple and can be built in any departmental shop. A 2-gal glass jug with a neck about 3 in. in diameter is used. A large stopcock is inserted in the center hole of the rubber stopper, and in the side holes are inserted a small stopcock connected to a short rubber tube, and the manometer tube. Since the accuracy of the results depends to a large extent on the sensitivity of the manometer, the latter is placed at an angle of about 45° and turpentine is used in it instead of water. The jug has an insulating cloth layer wrapped around it to increase the time required to establish thermal equilibrium.

The procedure is to blow air that is as dry as possible into the jug through the tube and small stopcock, to close the cock and allow the time necessary (about a minute) for the manometer reading to become steady. Reading h_1 is taken, and then the large stopcock is rotated once with a fairly fast motion to allow the air to reach atmospheric pressure by an adiabatic process. It then takes several minutes for the isovolumic change that results in the final manometer difference reading h_2. A typical set of readings is given in Table I.

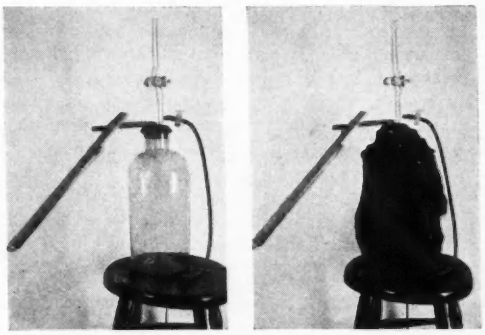


FIG. 1. Two views of the apparatus.

TABLE I. Typical set of data.

Trial	Manometer readings h_1 (cm)	h_2 (cm)	$h_1/(h_1-h_2)$
1	9.3	2.7	1.41
2	16.0	4.5	1.39
3	16.9	4.8	1.40
4	11.8	3.3	1.39
5	10.7	3.0	1.39
6	15.3	4.4	1.40
7	19.1	5.4	1.39
8	15.3	4.2	1.38
9	10.6	3.2	1.43
10	19.3	5.5	1.40
		Average	1.398

The present article merely calls attention to a simple apparatus that can easily be constructed and that gives amazingly good results as compared with more intricate and expensive forms. The only difference between this apparatus and that generally used is the use of a large stopcock instead of a poppet valve, and an inclined manometer for increased sensitivity. The use of the stopcock is definitely an advantage over a poppet valve since a poppet valve is hard to keep airtight when installed on a piece of apparatus as simple as this.

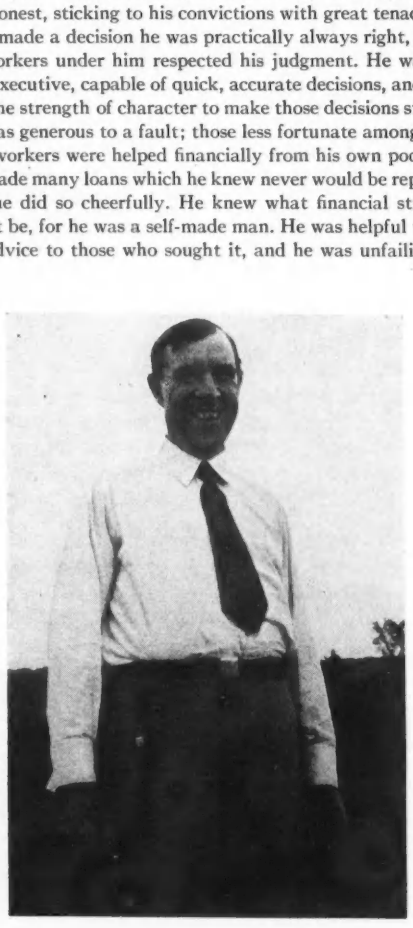
[1] J. de phys. (de la Metherie) **89**, 321, 428 (1819).
[2] For example, Cork, *Heat* (Wiley, ed. 2), p. 65.

F. K. Richtmyer: An Appreciation

FREDERICK R. HIRSH, JR.
University of Southern California, Los Angeles, California

PERUSAL of K. K. Darrow's "Third Richtmyer Memorial Lecture"[1] made me realize that F. K. Richtmyer's character has been described only by men who knew him as a fellow faculty member and counselor, while those who worked under him have not spoken. For this reason, and before time dims my memories, I wish to set down my impressions of him: a great man and a great "all-out" worker for the advancement of the profession and science of physics. I first saw him in 1926 at his annual lecture on x-rays in the course on recent topics of experimental physics given by Cornell staff members. I knew then that I had seen the man I wished to have direct my doctorate work; I never regretted the choice.

First, as a *man*, Richtmyer was utterly fearless, frank and honest, sticking to his convictions with great tenacity. If he made a decision he was practically always right, and we workers under him respected his judgment. He was a true executive, capable of quick, accurate decisions, and he had the strength of character to make those decisions stick. He was generous to a fault; those less fortunate among his own workers were helped financially from his own pocket. He made many loans which he knew never would be repaid, and he did so cheerfully. He knew what financial straits might be, for he was a self-made man. He was helpful with his advice to those who sought it, and he was unfailingly courteous and kind to his men. Seldom was he too busy to look in on his workers at their research. He knew that such visits kept up morale and he made them when and as he could. Last let me mention his gift of friendship: To those he knew well his friendship had no false restraints. He would come up to a friend or colleague fairly "wriggling" with joy (this is the only way to describe it). It was at just such a time, at Magog, Quebec, in 1932 on an eclipse expedition, that I snapped a camera picture of him, herewith reproduced: It was F. K. Richtmyer at his best, as those who knew him will agree. To sum up, there are few men of whom a person can say only good words—Richtmyer was one of them to me.



F. K. RICHTMYER

Next, as a *teacher*, Richtmyer was surpassed by few. I had the privilege of attending his lectures on modern physics before his book was written and could see things developing. These lectures were a joy to his students: He fairly burst with enthusiasm for his subject; his delivery was clear and faultless, for he knew whereof he spoke. He had a power of inspiring enthusiasm for his subject such as is given to few. One sensed instinctively that he was an inspiring expositor of the new physics.

Last, as a *physicist*, Richtmyer's greatest material contribution was his *Introduction to Modern Physics*. It summed up and correlated a theretofore unconnected and constantly growing mass of material, which up to the time of this critical summary had not been collected in one place and edited with a comprehending eye. His reputation could rest secure on this one work alone; but he also was a performer and leader in another field: His experimental work was the solid and accurate type which is the basis and essence of physics. He attended to each detail with care; experimental accuracy was his joy. He left no stone unturned in writing a paper; he investigated all possibilities; his clarity in writing was complete. His knowledge of physics was broad and comprehensive and inspired the confidence of his workers; "F. K." was never stumped for an answer in physics. I need not comment on his editorial activities, they are known by all. His role in the history of American physics is that of a truly great and inspiring leader in the new renaissance of physics. He was a man who believed in the place of physics as the King of Sciences, just as E. T. Bell has designated mathematics as the Queen of Sciences—and surely the two go hand in hand.

What was to be our last meeting occurred in the Athenaeum of the California Institute of Technology where, as Secretary of the Association of American Universities, Richtmyer was lunching with R. A. Millikan. He sprang up to greet me and say goodbye, as he was returning home. I never saw him again. His men returning to Cornell from the four corners of this country will greatly miss the warmth of his greeting and friendship. We who knew F. K. Richtmyer as one of his workers will never forget him; he set our first faltering footsteps in the pathway of science on the stepping stones of truth, inspired us, and sent us on our way rejoicing "as a strong man to run a race."

[1] Darrow, Am. J. Phys. **12**, 55 (1944).

DIGEST OF PERIODICAL LITERATURE

Training of Physicists for Chemical Industry*

In view of the real need in chemical and even biochemical industry for personnel trained in physics, it is desirable that more attention be given to the fitting of physics students for industrial posts. The properties that the physicist in industry will be called upon to measure include many that never have been properly defined. In treating such a situation the fundamental requirement is not the palliative of mere technical experience, which can never be progressive, but a more basic grasp of elementary physical principles. The outline that follows is a contribution to the structure of an appropriate training program.

The necessity for easy acquaintance with mathematical methods has always been obvious to the pure physicist. Yet a thorough treatment of industrial problems appears to demand mathematical knowledge of even a higher order than that needed in the academic field. Since the acquisition of such knowledge requires more time than the average physicist has available, it would seem desirable for the student to concentrate on those mechanical, geometrical and other short-cut methods developed for speeding up standardized mathematical operations, such as the use of arithmometers, slide rules and integrating machines. Work with the differential analyzer would not be out of place, as the adoption of this instrument in the larger industrial laboratories can reasonably be assumed to follow popular recognition of its usefulness. Nonmechanical devices for pancaking difficulties in advanced mathematics specifically concerned with experimental results of an arithmetic rather than an algebraic nature include the various graphical methods of presenting results and of solving equations, the use of special graph papers, line coordinate charts and nomography; and as well, the calculus of finite differences and the operational methods of solving differential equations.

In the university the physicist is accustomed to a careful preparation of the ground before performing an experiment. In industry this may be impossible and the would-be experimenter often finds himself reduced in role to an observer who, like the biologist or the astronomer, is partly in ignorance of the controlling factors at the time of his observations. Accordingly, the theory of errors and statistical methods in general should have an important place in his education. A prominent feature of control in chemical industry is its liability to change periodically, as, for example, with change of shift. However, the whole technic of weighting the measurements of a series of observers of different degrees of reliability is well known to physicists through attempts to find the most probable value of certain fundamental physical constants. This knowledge could with advantage be imparted to all who will enter industry either as research workers, trouble-shooters or process control chemists.

In dealing with phenomena too complicated for detailed analysis the method of dimensions is often of great help in clarifying the problem and in suggesting the most profitable mode of attack; it therefore is another suitable semimathematical topic for study by those destined for industrial posts.

In the study of the properties of matter, physics subdivides matter as solid or fluid in terms of its behavior or deformation. Chemical industry is more often concerned with the properties of a host of indefinite bodies intermediate between true liquids and true solids—gels, sols, suspensoids, plastics and similar bodies. Thus some treatment more complete than the simple subdivision of matter into solid, liquid and gas is needed for one who will deal with industrial materials. This treatment should cover phenomena beyond the limit of stress at which ideal behavior ceases to be a reasonable approximation—such phenomena as ultimate strength, elastic hysteresis, work-hardening, fatigue, the relaxation time, plasticity, thixotropy, dilatency and other rheological properties.

The treatment of motion must be based on the fact that industry is much less concerned with rotations and translations than with the indefinite type of movement typified by the various diffusive mechanisms. The erratic path of a random motion, such as the Brownian movement, is characterized by a factor of dimensions $L^{-2}T^{-1}$, and a factor of similar dimensions is at the basis of all diffusive processes. The essential difference between activated and true diffusion should be pointed out; so also should be the fact that one material carrier can transport either mass, energy, momentum or electric charge. The interrelations among mass diffusion, viscosity, and thermal and electrical conductivity need attention, particularly as the physical chemist appears to be trained to look upon diffusion and viscosity as opposing properties, instead of parallel manifestations of a single transport phenomenon. The parallelism between momentum transfer and energy transfer is comparatively well known in the treatment of heat-flow problems; but the allied relationship between mass transfer and energy transfer is less well known, though it is of even greater importance in chemical engineering, particularly in all those processes in which mass and heat are simultaneously interchanged, as in fractionating columns and spray towers.

Transfer of heat by convection and the analogous properties of turbulent resistance and turbulent mixing of fluids is a subject to which little attention ordinarily is given in physics courses; yet it is encountered on every side in industry.

Finally, the physicist for chemical industry needs a thorough grounding in all scientific instruments, particularly in those used or of potential use in the measurement of the properties employed as process variables. This grounding should be coupled with a knowledge of the principles of measurement and of control, and with enough insight into the significance of the properties to be measured to enable one to picture to himself the meaning of his

measurements and to foresee the consequence of any departure, either on his own part, or through changes in the instrument, from the standard method of measurement. —R. C. L. BOSWORTH, *J. Sci. Inst.* **20**, 142–145 (1943).

* The original article also contains a section on specific ways in which the physicist, by virtue of his special knowledge and training, can help in problems peculiar to chemical industry.

For a Sane Approach to Tomorrow's World

Science with its technology—unquestionably the brilliant tool to nature's munificence and a high tribute to the genius of man—is not an Aladdin's lamp, to be rubbed at every whim. The war has shown clearly the double-edged sword it really is. Less obvious are the disappointments, the inner pain, the moral disarray that flow from a technology that tempts mankind with goods that master him. What matters it if a man has a high speed car and only a roadhouse to go to; or if he can talk to Chungking and has nothing to say? Warmth and shelter, varied food and protected health, the implements of education and social understanding, relief from pain and freedom from drudgery —insofar as science supplies these, we bless it; they are what gives man the opportunity to go forward to the perfection of which we in the democratic tradition believe him capable. But the things that clutter, that burden under pretense of lifting, these we damn eternally.

Let us approach tomorrow's world with caution. Let us be sure that the plastics are moulded to forms that have meaning, that the alloys are cast into shapes that possess the true grace of honest purpose. We need not be Thoreaus and toss away paperweights which only collect dust, but we can learn from him the lesson of freedom from possessions and the joys of simple living.

An urban society exacts enormous sacrifices of its members. It gives them many concentrates that are good. But often the price is more than can be paid, and bankruptcy follows in the form of neuroses and mental strains. Will the men who plan the products of tomorrow consider this? Whatever the device or gadget, it must be contemplated in terms of the *whole* environment. It is not enough that the prefabricated house will keep us warm in winter, or cool in summer. It must as well make concessions to the folkways and taboos of our culture and groups.

An enormous capacity to produce will be available at the end of this war, and the drive to find new goods to utilize it will be enormous. We want this used to destroy poverty and crime, to light the dark corners of our social system, to give us all the things we need physiologically and sociologically. But we definitely don't want it for frustrations.

Let us weigh carefully what suits us best. Used carelessly, our great machines can destroy us. Used wisely, we will achieve an age that even Pericles could not match.—WILLIAM S. LYNCH, Department of Humanities, The Cooper Union, in *Think* **10**, 5 (1944).

Portraits of Book-Reviewers Drawn by Themselves

A just review is a difficult, rare and highly praiseworthy achievement.

Readers of a serious book review may rightfully expect to find in it two portraits, one of the book and one of the reviewer. A reviewer may fail to portray the book but he cannot fail to portray himself—he is pictured by his performance.

If he portrays the book, he thereby portrays himself as having intellectual and scholarly competence and an imperious sense of honor including loyalty to the author, to the editor, to the public and to truth. His picture is that of a worthy citizen of the commonwealth of science and letters.

If he fails to portray the book, he thereby portrays himself as one lacking intellectual or scholarly or moral competence or two of these or all three of them. All such portraits are spiritually ugly. Of all of them the ugliest is perhaps that of a reviewer who uses the book merely or mainly as a trapeze upon which to mount and display himself. It is the picture of one who is vain, deceitful, and cowardly—intellectually a knave, morally a fool.

A periodical "Review of book reviews" could render a very great and precious service. Its chief function would be, on the one hand, to signalize and commend competent reviews and reviewers, and, on the other, to signalize and denounce incompetent reviews and the nasty little scoundrels who perpetrate them.—CASSIUS JACKSON KEYSER, *Mole philosophy and other essays*, quoted by *Scripta Mathematica* **9**, 184 (1943).

Check List of Periodical Literature

Science and mathematics in educational programs for returning service men and women. Cooperative Committee on Science Teaching, *Sch. Sci. and Math.* **44**, 517–520 (1944). This particular report deals with the training, for a relatively short period of about one year, of returning service personnel who desire instruction below the college level. Some idea of the size and complexity of the postwar educational problem is gained from a recent study made by the Army Service Forces; it indicates that white enlisted men in the United States have the following postwar plans for returning to school or college; 7 percent will return full time regardless of government aid and more than half of these even if offered a good job; 28 percent will return full time if given government aid; 17 percent expect to return part time.

Newtonian and other forms of gravitational theory. G. D. Birkhoff, *Sci. Mo.* **58**, 49–57, 135–140 (1944). An excellent survey showing the intimate formal relation between Newton's theory of universal gravitation and certain proposed relativistic modifications. The Newtonian theory doubtless will stand as the realistic basis for astronomical calculations, with relativistic theories used only in a few cases involving high velocities. Nevertheless, the latter theories seem more in accord with the electromagnetic structure of matter and, since they are as yet in a highly incomplete state, deserve much more serious attention than they have hitherto received.

Science without experiment: a study of Descartes. R. Suter, *Sci. Mo.* **58**, 265–268 (1944).